Université Paul Cézanne Aix-Marseille III
École Doctorale Physique et Sciences de la Matière
Laboratoire d'Astrophysique de Marseille

Masques Photométriques et Détection des transits planétaires

Dans le cadre de la mission Corot

THÈSE

pour obtenir le grade de

Docteur de l'Université Paul Cézanne

Discipline : Optique, Image et Signal

par

Pascal Guterman

Soutenue publiquement le *29 novembre 2005* au
Laboratoire d'Astrophysique de Marseille

Devant le Jury composé de :

Pr.	Annie Baglin	Examinatrice
Pr.	Salah Bourennane	Examinateur
Dr.	Antoine Llebaria	Directeur
Dr.	Pierre Barge	Directeur
Pr.	David Naccache	Rapporteur
Dr.	Fabio Favata	Rapporteur
Pr.	Mustapha Ouladsine	Président

2005 AIX3-0041

Université Paul Cézanne Aix-Marseille III
École Doctorale Physique et Sciences de la Matière
Laboratoire d'Astrophysique de Marseille

Masques Photométriques et Détection des transits planétaires

Dans le cadre de la mission Corot

THÈSE

pour obtenir le grade de

Docteur de l'Université Paul Cézanne

Discipline : Optique, Image et Signal

par

Pascal Guterman

Soutenue publiquement le *29 novembre 2005* au
Laboratoire d'Astrophysique de Marseille

Devant le Jury composé de :

Pr.	Annie Baglin	Examinatrice
Pr.	Salah Bourennane	Examinateur
Dr.	Antoine Llebaria	Directeur
Dr.	Pierre Barge	Directeur
Pr.	David Naccache	Rapporteur
Dr.	Fabio Favata	Rapporteur
Pr.	Mustapha Ouladsine	Président

2005 AIX3-0041

Remerciements

Cette thèse au Laboratoire d'Astrophysique de Marseille fut initiée dans le cadre d'une collaboration avec la société Gemplus qui m'emploie. Elle m'a donné la chance de faire connaissance avec la communauté scientifique par le biais de la mission Corot . Le privilège de cette expérience a dépassé de loin toutes mes espérances et le plaisir de travailler au LAM fut quotidien.

Le souvenir de ces années passionnantes et intenses reste si vivace qu'on dirait que tout s'est déroulé dans la seule journée d'hier. Le contact avec le monde astronomique était un rêve d'enfant. Je ne remercierais jamais assez David Naccache, scientifique et dirigeant d'exception à Gemplus, pour sa compréhension et sa flexibilité. C'est à lui que je dois la réalisation de ce rêve. Je le remercie également d'avoir accepté la charge de rapporteur.
La plus importante des rencontres qui ont enrichi ce parcours fut sans conteste celle d'Antoine Llebaria, responsable du traitement d'image au LAM et Directeur de cette thèse. Je tiens à exprimer tout particulièrement ma gratitude et mon amitié envers cet esprit aussi brillant que sympathique. Il a toujours trouvé du temps à me consacrer et je lui dois la plus grande partie de ce que j'ai appris.
Je remercie chaleureusement Pierre Barge, également Directeur de cette thèse. Il m'a guidé tout le long du trajet et son soutien humain de tous les instants m'a été crucial. Il m'a "éveillé" à l'astronomie et passionné au sujet de la formation planétaire. Je le remercie aussi pour le temps considérable consacré, pour ses encouragements à la publication et son travail de relecture.
Je suis particulièrement honoré et touché qu'Annie Baglin ait eu la gentillesse de faire partie du jury malgré ses énormes responsabilités de PI Corot . Je suis heureux d'avoir fait partie de son équipe et lui exprime toute mon admiration.
Je tiens également à remercier Salah Bourennane et Mustapha Ouladsine qui ont aimablement accepté la charge de membre et président du jury. Je remercie également Fabio Favata pour son travail de rapporteur.

Raphael Cautain m'a beaucoup aidé par sa culture absolue de Corot et la clarté de ses explications. Je suis redevable à Céline Quentin qui m'a aidé pour la publication sur la détection des transits. Je sais gré à Eric Brier pour son aide dans les calculs.
Nombre de personnes m'ont été d'un grand secours et il serait impossible de les énumérer toutes. Je remercie entre autres Jean-Charles Meunier pour son aide, Thomas Fenouillet qui m'a presque persuadé que l'ordinateur était un objet docile, Claire Martin en souvenir de nos fins de thèses respectives lorsque 3

heures du matin était une heure ordinaire pour communiquer, les personnes de l'administration et les étudiants dont l'entraide m'a épargné nombre d'errements.

Au cours de cet itinéraire, Magali Deleuil ainsi que Claire Moutou m'ont ouvert la porte d'une mission d'observation à La Palma pour prendre un peu de recul. Même si c'était une nuit de 31 décembre et qu'il a fallu combiner 34 vols bons marchés (à quatre) afin de rester en famille, je les remercie pour cette palpitante épopée. Message : je reste volontaire. La palme de la bonne humeur revient sans conteste à Magali, ainsi que tout le palmier.

Au-delà de ce travail de thèse, je tiens à rendre grâce à Claire, ainsi que Jean-Claude Bouret, Stéphane Arnouts, François Bouchy, Laurence Tresse, Stéphane Basa, Christophe Adami et tant d'autres astronomes qui ont répondu avec patience et pédagogie au flot de mes questions sur leur extraordinaire métier.

J'adresse toute mon affection à mes parents, et remercie en particulier ma mère à qui j'ai donné quelques soucis à l'époque de mon orientation scolaire, même si ce n'est qu'après l'équivalent de 20 ans de redoublements.

Enfin, Joëlle (Madame) fut le papillon de "l'effet papillon" : quand elle eut l'idée fondatrice "Pourquoi ne reprendrais-tu pas les études ?", elle ne pouvait soupçonner l'ampleur du basculement climatique qui se déclencherait en retour. En effet, bien que déjà salarié d'une entreprise, je m'engouffrais aussitôt dans la passion de cette deuxième vie en parallèle. Du DEA jusqu'en fin de thèse, les conséquences pour elle et pour nos filles Camille et Julie furent parfois pesantes. Elle a pourtant assumé, résisté à la tentation légitime de sortir mes affaires sur le perron, et a toujours continué de m'encourager.

Je te dédie, Joëlle, ces années et cette thèse.

Table des matières

1 Introduction générale **3**

2 La détection des planètes extra-solaires **9**
2.1 Les différentes méthodes 9
2.1.1 Vélocimétrie radiale 9
2.1.2 Astrométrie 11
2.1.3 Méthode par réflexion 12
2.1.4 Observation directe 12
2.1.5 Lentilles gravitationnelles 12
2.2 la Méthode des transits 13
2.2.1 Probabilité 14
2.2.2 Durée du transit 15
2.2.3 Intérêt du spatial 16
2.2.4 Le bruit photonique 16
2.2.5 La variabilité stellaire 17

3 La recherche de planètes extrasolaires avec Corot **23**
3.1 Présentation de la mission 23
3.2 l'Instrument 26
3.2.1 Le baffle 26
3.2.2 La défocalisation de l'image 28
3.2.3 Le système disperseur 29
3.2.4 Les PSFs de référence 29
3.2.5 Position de référence 31
3.3 Photométrie 31
3.3.1 Photométrie par ajustement de PSF 31
3.3.2 Photométrie d'ouverture 33
3.3.3 Sélection des cibles 34
3.4 Capacité de détection 35
3.4.1 Nombre de détections attendues 37

I Fenêtrage optimisé **39**

4 Énoncé des contraintes pour l'optimisation des masques photométriques **41**
4.1 Terminologie 42
4.2 Critère de qualité pour le fenêtrage 43

5 Calcul des masques optimaux **45**
5.1 Le signal 45
5.2 Les bruits 46
5.2.1 Le bruit photonique 46
5.2.2 La contamination 46
5.2.3 Le jitter 47
5.2.4 La respiration 48
5.2.5 Les bruits électroniques 49
5.2.6 Le traînage 49
5.2.7 La saturation 50
5.2.8 Bruits d'arrondi 50
5.2.9 La variabilité stellaire 51
5.2.10 Les éclipses d'étoiles du fond 51
5.3 Modèle du bruit photonique 52
5.4 Modèle jitter 1-D 53
5.4.1 Signal 53
5.4.2 Bruit au 1^{er} ordre 54
5.4.3 Bruit au $2^{ème}$ ordre 56
5.5 Modèle jitter 2-D 56
5.5.1 Bruit 2-D au 1^{er} ordre 57
5.5.2 Bruit 2-D au $2^{ème}$ ordre 58
5.6 Vérification expérimentale 59
5.7 Simulation d'Images 62
5.7.1 Simulation des PSFs stellaires 62
5.7.2 Masques optimaux 64
5.8 Première publication Llebaria et al. (2002), SPIE. 66

6 Réduction Optimisée du nombre de patrons **67**
6.1 Introduction au problème de la réduction optimisée 67
6.2 Nécessité et contraintes de la réduction 68
6.3 Méthode 1 : Paramétrisation *a priori* 69
6.4 Méthode 2 : Paramétrisation *a posteriori* 70
6.4.1 Pertinence de la réduction : 70
6.4.2 Dimension sous-jacente 71
6.4.3 Formalisation de la méthode utilisée précédemment 72

6.4.4 Défauts de cette approche 75
6.5 Méthode 3 : L'homogénéisation morphologique directe 75
6.5.1 Algorithme de Base : la Condensation 76
6.6 Méthode 4 : Le problème à K-moyennes 78
6.7 Méthode 5 : Le tri efficace de masques pseudo-aléatoires 81
6.7.1 Dilemme de l'optimisation collective 82
6.7.2 Mesure du S/B global 84
6.7.3 Tolérance aux masques aléatoires 84
6.7.4 Description de la méthode 86
6.7.5 Résultats obtenus 87
6.7.6 Mise en oeuvre de la procédure 89
6.7.7 Conclusion 90
6.8 Deuxième publication Llebaria et al. (2003), SPIE. 93

7 Attribution des patrons sur une image complète 95
7.1 Méthodes testées 96
7.2 Méthode choisie 97
7.3 Résultat 98
7.4 Double critère de priorité 100
7.5 Données destinées à la détection des transits 104
7.6 Conclusion et perspectives pour le fenêtrage 107

II Détection des transits dans les courbes de lumière 109

8 Énoncé des contraintes. 111
8.1 Introduction 112
8.2 Les méthodes de détection 112
8.3 Test en aveugle 116
8.3.1 Les données de simulation 116
8.3.2 Les méthodes utilisées 118
8.3.3 Traitement des données manquantes 120

9 La méthode proposée 123
9.1 La détection 124
9.1.1 Définition du débruitage et de la détection 124
9.1.2 Justification du séquencement inversé 124
9.1.3 Détecteur utilisé 126
9.1.4 Modèles de transit 128
9.1.5 Gabarit optimal 129
9.1.6 Premier résultat 129
9.2 Le débruitage 131

9.2.1 Prépondérance des bruits systématiques 131
9.2.2 Identification des modes communs 132
9.2.3 Obtention des courbes de vraisemblance 134
9.2.4 Classification des courbes de lumière 135
9.2.5 Caractérisation des événements 137
9.3 Résultats 138
9.4 Perspectives 141
9.5 Troisième publication Guterman et al. (2005), SPIE. 142

10 Conclusion **143**
10.1 Acquisition des données 143
10.2 Détection des transits 145

Annexes

A Performances des patrons **149**

B Le test en aveugle **151**
B.1 Article de Moutou et al. (2005), A&A. 152

Résumé

La mission spatiale Corot sera lancée fin 2006. L'un de ses objectifs est de détecter des planètes extrasolaires dont la taille n'excède pas quelques rayons Terrestre. Corot utilise la méthode des transits. Il s'agit de déceler, à l'aide d'un petit télescope, la très faible baisse de flux d'une étoile causée par le passage d'une planète devant le disque stellaire. Pour réaliser son objectif, Corot atteindra une précision photométrique de 7.10^{-4} sur des étoiles de magnitude 15.5 intégrées pendant une heure. Corot mesurera ainsi 60 000 étoiles sur l'ensemble de sa mission.
Les étoiles sont mesurées simultanément et continûment par champs de 12 000 cibles grâce à l'utilisation d'une matrice CCD. Pour limiter le volume de données, Corot réalise à bord une photométrie d'ouverture : à chaque pose, les pixels compris dans une ouverture entourant chaque étoile-cible sont sommés en une mesure photométrique unique. Idéalement, chacune des 12 000 fenêtres est optimisée en fonction des paramètres et de l'environnement de son étoile-cible. Le problème est qu'il faut réduire la variété de forme des ouvertures (les "patrons") d'un facteur 20, pour rester compatible avec la capacité de traitement des données à bord du satellite. Il s'ensuit une perte de spécialisation des fenêtres, contradictoire avec le maintien d'un rapport signal à bruit maximal pour toutes les cibles. Le centre de la question est donc de déterminer une collection réduite de patrons qui satisfasse l'objectif scientifique de Corot . C'est un problème très important pour le succès de la mission car la qualité des mesures de Corot dépend directement du choix des fenêtres. Un deuxième problème important est la détection des transits planétaires dans les multiples courbes de lumière temporelles que produira Corot .
La première partie des travaux que j'ai menés est consacrée à l'étude des méthodes de réduction des patrons, et à leur répartition sur les étoiles cibles en préservant au mieux le rapport signal à bruit. La deuxième partie de l'étude s'attache à la détection des transits dans les courbes de lumière temporelles que produira Corot . L'organisation en deux parties de ce mémoire traduit ce double thème.

Les chapitres introductifs 1 et 2 esquissent le contexte et exposent les méthodes de détection des planètes extrasolaires en détaillant la méthode des transits plané-

taires. Le chapitre 3 dresse une description du volet exoplanètes de Corot et en précise l'objectif.
La première partie du manuscrit est consacrée à l'optimisation des fenêtres. Le chapitre 4 situe le problème. Mon travail a commencé par l'étude d'une expression analytique pour le rapport signal à bruit (chapitre 5). Puis, la résolution du problème se déroule en 3 étapes :

- L'obtention de 12 000 masques de lecture, chacun optimisé pour sa cible, sert de base pour la réduction.
- Le chapitre 6 relate l'étude sur la réduction de diversité de ces masques. J'ai testé différentes méthodes de classification, ou de regroupement suivant des critères morphologiques. La méthode mise au point et finalement retenue est un tri efficace de masques pseudo aléatoires.
- Au chapitre 7, j'ai comparé des méthodes d'assignation des patrons aux cibles. Le problème est d'attribuer les patrons en gérant des priorités pour éviter les chevauchements entre fenêtres. Ces chevauchements sont causés par l'encombrement des champs stellaires visés. La technique mise au point est basée sur une gestion de file d'attente.

La deuxième partie du mémoire est consacrée à la détection des transits planétaires dans les courbes de lumière de Corot . Leur signature est une baisse périodique de luminosité de forme caractéristique. Les transits de plus faible amplitude et de plus grande période correspondent aux planètes recherchées, mais sont aussi les plus difficiles à détecter. L'amplitude des transits détectables est limitée par le bruit en aval du détecteur ; les pics de détection correspondant aux petits transits tendent à rester noyés dans le continuum de détection.
J'ai étudié la façon d'améliorer la détection en atténuant dans le continuum les artefacts liés à des facteurs systématiques. Pour identifier ces artefacts, la méthode proposée tire parti du grand nombre des courbes de lumière qui seront produites par Corot . Le chapitre 8 expose le problème, dresse un panorama des méthodes de détection et décrit l'élimination des bruits dans une collection de courbes simulées. Le chapitre 9 décrit la méthode proposée, basée sur la pondération des facteurs systématiques présents dans les courbes de corrélation temporelle.

La conclusion générale est présentée au chapitre 10. Les méthodes proposées pour la réduction des patrons et leur affectation aux cibles permettent de préserver le plus souvent 95% à 97% du rapport signal à bruit initial. Une étoile est d'autant mieux préservée qu'elle est brillante. Ces techniques seront utilisées pour déterminer les masques de vol. Pour la détection des transits, le recours à l'identification et au traitement des systématiques a permis de réduire significativement le bruit dans la détection, faisant émerger des transits supplémentaires.

Chapitre 1

Introduction générale

Le premier but du travail présenté dans cette thèse est de déterminer les ouvertures photométriques optimales servant à fenêtrer l'image des étoiles sur la matrice CCD embarquées à bord de satellites conçus pour la recherche des transits exoplanétaires.
Le second but de ce travail est de présenter l'utilisation qui peut être faite du grand nombre des étoiles cibles d'un champ d'observation pour améliorer la capacité de détection des signaux de transits dans les courbes de lumière.
Le travail réalisé s'inscrit dans le cadre de la préparation de la mission spatiale Corot , ses différentes parties ont fait l'objet de 4 publications figurant dans ce manuscrit.

Depuis la prise de conscience que les myriades d'étoiles qui emplissent notre ciel nocturne sont autant de Soleils, l'hypothèse de planètes gravitant autour de l'une d'elles devînt inéluctable. C'est Mayor & Queloz (1995) qui ont détecté pour la première fois l'un de ces mondes autour de l'étoile 51-Pegasi. Depuis, les découvertes se succèdent et s'accélèrent. Après maintenant une décennie, on connaît 168 planètes extra-solaires, ou exoplanètes, et 18 systèmes multi-planétaires[1].

Le nombre d'objets détectés est suffisant pour aborder les faits du point de vue statistique. La table 1.1 de masse des exoplanètes montre une abondance accrue de planètes peu massives, ce qui laisse espérer nombre de planètes telluriques de quelques masses terrestres. Vers les faibles masses, la coupure marque la limite de sensibilité. A l'autre extrémité la raréfaction au-delà de 10 M_J, nommée "désert des naines brunes" annonce la transition vers une autre classe d'objets.

[1]Une encyclopédie électronique est maintenue à jour par l'observatoire de Meudon au lien `http ://www.obspm.fr/planets`

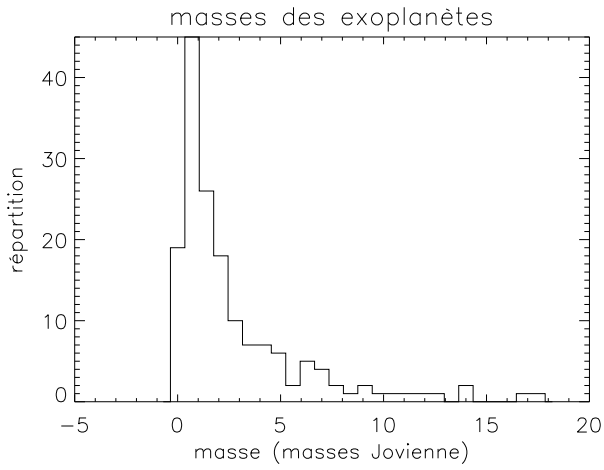


FIG. 1.1 – *Diagramme des masses d'exoplanètes, en masse Jovienne. La diminution de population vers 10 M_J n'est pas un biais de mesure car ces astres sont les plus facilement détectables. Pour comparaison la masse Terrestre $M_\oplus \simeq 0.003 M_J$*

D'autres statistiques, telle la distribution comparée entre la masse et la période, ou la probabilité de présence d'une planète en fonction du taux d'éléments lourds de l'étoile permettent de contraindre les mécanismes de formations planétaires. La corrélation avec la métallicité montre par exemple qu'étoiles et planètes se forment suivant des scénarii très différents.

Les principaux mécanismes proposés mettent en jeu aussi bien les forces gravitationnelles, radiatives ou magnétiques que les forces de friction aérodynamiques. Il semble aujourd'hui que les grandes lignes d'un scénario de formation planétaire se dessinent. A l'issue de la formation de l'étoile, il subsiste un disque de gaz et de poussières : le disque protoplanétaire. Les poussières vont se coller sous l'effet de l'attraction électrostatique et moléculaire, formant de minuscules grains de $1\mu m$ à 1 mm (coagulation). Ces grains s'agglomèrent ensuite, tout en étant contrariés par les fréquents chocs à haute vitesse relative qui tendent à les désagréger. Au-delà d'une taille critique, la matière rencontrée s'accumule sans dommage, formant des planétésimaux. Cette étape rapide se déroule en moins de 10^5 années, les corps sont alors assez massifs pour que le gaz ne joue plus de rôle dans leur mouvement. Certains corps massifs dans les régions externes du disque pourraient capturer une grande quantité de gaz environnant en quelques millions d'années, devenant ainsi des planètes géantes gazeuses. La totalité du processus doit avoir pris place avant que le gaz ait disparu, soufflé par l'étoile. Dans la partie plus interne du disque, les collisions entre corps solides formeraient des planètes rocheuses en 10 à 100 millions d'années. Ces collisions expliqueraient la

cratérisation observée des planètes actuelles. Le temps de formation peut être plus long car le gaz n'y joue plus de rôle.

L'étape de formation planétaire est complexe et il se pourrait que des tourbillons à grandes échelles se forment dans le disque protoplanétaire, aidant le matériau solide à se concentrer et s'agglomérer pour former des corps de taille astéroïdale. Barge & Sommeria (1995) proposent un mécanisme qui repose sur la persistance de tourbillons stables à grande échelle dans le gaz de la nébuleuse protoplanétaire. Un anticyclone persistant capturerait un grand nombre de particules, les forçant à s'agréger en planétésimaux. La masse capturée présente un maximum à hauteur de l'orbite de Jupiter. Au-delà de cette distance, les matériaux capturés de faible densité sombreraient profondément dans le vortex en s'effondrant pour donner naissance au noyau solide des planètes géantes. En deçà de cette orbite, la capture sélectionne plutôt les particules denses, les agrégeant en un planétésimal assez massif pour être éjecté du maëlstrom. Il fusionnerait alors avec d'autres pour former une planète tellurique.

Loin d'être un simple comptage, la quête des exoplanètes révèle un trésor de diversité. La table 1.1 présente quelques cas remarquables. La plupart sont des planètes massives gravitant près de leur étoile. Ces "Jupiter chauds" peuvent atteindre des températures de l'ordre du millier de degré. Les modèles de formation planétaires situaient jusqu'alors les géantes gazeuses loin de l'étoile pour y trouver suffisamment d'atomes légers. Les durées de révolution s'étalent entre 1.2 jours et 12 ans. Vers les faibles masses, *Mu Area d (HD 160691d)* fut la première candidate planète de masse comparable à celle de Neptune tellurique avec moins de $10M_{\oplus}$ (masses terrestre). Dans un autre registre, on trouve des systèmes multi-planétaires, mais aussi de plus insolites systèmes multi-stellaires où le record appartient à *HD 188753A b* qui gravite autour d'une des composantes d'un système triple. On s'explique encore mal comment une planète a pu se former dans un environnement aussi secoué de perturbations gravitationnelles.

Cette abondance de cas atypiques a bouleversé notre vision de la formation des systèmes planétaires. On pensait qu'à l'image de notre propre système, la formation planétaire produisait de petites planètes denses et proches de l'étoile, et des géantes gazeuses éloignées. Les premières agglomèrent les atomes lourds résiduels de la formation de l'étoile, les secondes trouvent assez d'atomes légers dans la large circonférence du disque circumstellaire. Les planètes auraient gravité sur le lieu même de leur formation. Les Jupiter chauds semblent être en contradiction avec cette idée. Ils sont parfois multiples et paraissent trop proches de l'étoile pour avoir trouvé sur place la matière suffisante à leur formation. On envisage à présent un mécanisme de migration qui opère alors que la planète

TAB. 1.1 – *Exoplanètes remarquables*

Exoplanète	Particularité	masse (M_J)	orbite (jour)	date
51 Pegasi (HD 217014) b	La première	0.468	4.2	(Mayor & Queloz 1995)
HD 202206 b	La plus massive	17.4	255	
Gliese 876 (HIP 113020) d	La tellurique la moins massive	7.3 $M_{\mathrm{T}erre}$	1.9	2005
OGLE-TR-56 b	L'orbite la plus brève	1.45	1.21	(Torres et al. 2003)
Osiris (HD209458) b	Détection d'Oxygène et Carbonne atmosphériques à 1 000 °C s'accrétant sur l'étoile	0.69	3.5	(Henry et al. 1999), (Vidal-Madjar et al. 2004)
Upsilon Andromedae (HD9826) b,c,d	Le premier système stellaire	0.69 1.89 3.7	4.6 241 1284	(Marcy & Butler 1996)
55 Cancri (HD 75732) b,c,d,e	Système à 4 exoplanétes, avec ceinture de Kuiper et orbite la plus longue	0.784 0.217 3.92 0.045	14.7 43.9 4517 2.81	1996 à 2004
Mu Area (HD 160691) d	La première tellurique	14 $M_{\mathrm{T}erre}$	9.55	2004
PSR :B1257+12	Exoplanète autour d'un pulsar			(Wolszczan 1994)
2M1207	La première image directe. Exoplanète autour d'une naine brune	5	2450 ans	(Chauvin et al. 2004)
HD 188753A b	Exoplanète autour d'une des étoiles triples HD 188753A,B,C	1.14	3.34	(Konacki 2005)

à presque terminé son accrétion. Avec le temps, la planète se rapprocherait de l'étoile et pourrait dans certains cas y sombrer. A la lumière de ces découvertes, notre système Solaire n'est plus la règle mais fait plutôt figure d'exception.

L'autre source d'intérêt majeur des exoplanètes est bien entendu la quête philosophique et scientifique d'autres éclosions de la vie. Dans ce sens, Vidal-Madjar et al. (2004) ont détecté autour de la planète *Osiris* la présence de Carbone et d'Oxygène dans la haute atmosphère, certes en train de tomber sur l'étoile. Il sera nécessaire de repousser toujours plus loin les limites de la détection d'exoplanètes afin de collecter des informations les plus précises sur la nature et la composition de ces objets. Quelle que soient les spéculations exobiologiques, les planètes-hôte sont recherchées dans la zone "habitable" d'une étoile, située à une distance de l'ordre de l'unité astronomique (UA). On adopte comme critère d'habitabilité la plage de pression et de température compatible avec la présence d'eau liquide. Les planètes trop proches ne sont pas compatibles avec une chimie élaborée, l'excès de chaleur et de rayonnements ultraviolet brise les assemblages moléculaires. Les mondes éloignés sont froids et figés.

A part le cas 2M1207 qui est une étoile naine rouge peu lumineuse, les exoplanètes ne peuvent pas être observées directement à ce jour. Elles sont un milliard de fois moins lumineuses que leur étoile, et trop proches d'elle au regard de la distance qui nous en sépare. On a donc recours à des méthodes indirectes qui décèlent l'effet des perturbations gravitationnelles ou lumineuses que la planète induit sur l'étoile. La détection est d'autant plus difficile que les planètes sont petites et éloignées de leur étoile. C'est pourquoi une précision accrue est nécessaire pour détecter et étudier les planètes "habitables". Schneider (1999) dresse un tour d'horizon des méthodes et projets. En attendant la réalisation d'interféromètres spatiaux ambitieux, les techniques les plus appropriées sont la vélocimétrie radiale et la méthode des occultations dite *méthode des transits*. La vélocimétrie radiale donne accès à la masse et à la période de la planète, donc au demi grand axe de son orbite car la masse de l'étoile peut être estimée de façon assez précise. Les transits indiquent la taille de cette planète et affinent la détermination de la période. Ces deux méthodes sont complémentaires et la combinaison de leurs informations permet de déterminer complètement le système, de lever les ambiguïtés, et de déduire la densité de la planète donc son éventuelle nature rocheuse. Elles se prêtent toutes deux à des études systématiques.

Chapitre 2

La détection des planètes extra-solaires

Ce chapitre présente les différentes méthodes de détection des exoplanètes et détaille les principaux perturbateurs intervenant dans la méthode des transits.

2.1 Les différentes méthodes

2.1.1 Vélocimétrie radiale

Cette méthode spectroscopique est à l'origine de la quasi-totalité des découvertes ou confirmations à ce jour. Dans un système planète-étoile, l'attraction réciproque fait légèrement tourner l'étoile autour du centre de masse commun. Vue de la Terre, l'étoile oscille autour de sa position d'équilibre (voir Fig.2.1). La composante de ce "va-et-vient" vers l'observateur provoque un décalage Doppler récurrent du spectre. Une spectroscopie précise permet de remonter à $v_\star(t)$, vitesse de l'étoile à différentes dates.

A partir de $v_\star$ et T, période du mouvement, on cherche à retrouver la masse m de la planète. A tout moment, les vitesses de l'étoile et de la planète sont liées par la dérivée de la relation barycentrique :

$$v_\star = v\frac{m}{M+m} \tag{2.1}$$

où v est la vitesse de la planète, M la masse de l'étoile connue par le type de son spectre. On voit que plus la planète est proche de son étoile, plus elle est rapide et donc détectable. Dans le temps T, elle parcourt une orbite de circonférence $2\pi a$ où a étant le demi grand axe de la planète, donc $v = 2\pi a/T^3$ (on néglige le mouvement de l'étoile). Cette relation nous permet d'éliminer a en utilisant la $3^{ème}$ loi de Kepler et de relier v à M :

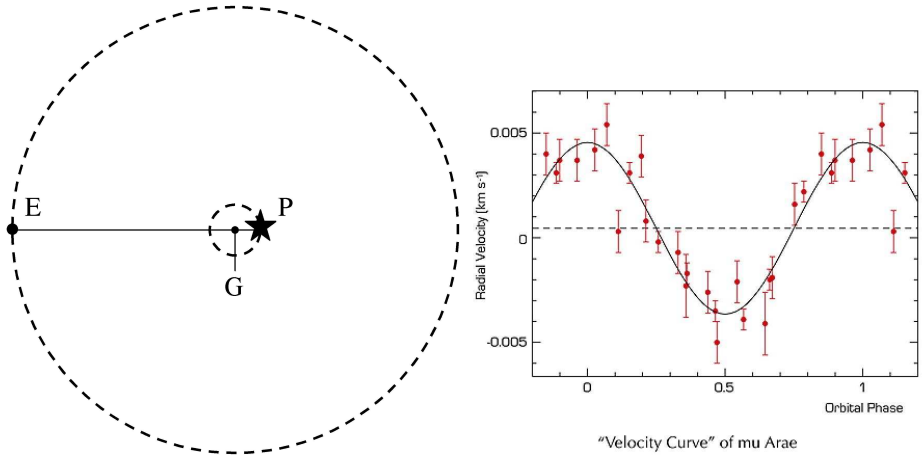


FIG. 2.1 – *A gauche, trajectoire d'un système planète-étoile autour du barycentre commun G. A droite, vitesse de Mu Area mesurée à différents instants par le spectromètre HARPS. La figure montre le meilleur ajustement par une sinusoïde, la courbe étant repliée à la période estimée de la planète.*

$$\frac{\mathcal{G}(M+m)}{4\pi^2} = \frac{a^3}{T^2} \tag{2.2}$$

$$= \frac{1}{T^2}.\frac{v^3T^3}{(2\pi)^3} \tag{2.3}$$

$\mathcal{G}$ est la constante de gravitation universelle. En revenant à $v_\star$ (Eq.2.1) :

$$v_\star = m\left(\frac{2\pi\mathcal{G}}{T(M+m)^2}\right)^{\frac{1}{3}} \tag{2.4}$$

Comme l'effet Doppler ne donne accès qu'à la composante radiale $v_{\star r}$, il faut multiplier les termes par $\sin i$, l'inclinaison de l'orbite sur la voûte céleste. Finalement :

$$m\sin i = v_{\star r}\left(\frac{TM^2}{2\pi\mathcal{G}}\right)^{\frac{1}{3}}$$

projetée sur la ligne de visée La masse de la planète n'est connue qu'à un facteur sinus près. Mais cela influe peu sur l'ordre de grandeur de m car au sens d'une distribution uniforme 66% des angles ont un sinus supérieur à 0.5.

La résolution de vitesse en vélocimétrie radiale atteint 3m.s^{-1}, ce qui paraît prodigieux concernant un astre étendu, couvert de cellules convectives opérant sur des dimensions de l'ordre du millier de kilomètres, de surcroît en rotation

différentielle, et animé d'une vitesse propre de plusieurs km.s^{-1}, éjectant de la matière par intermittence, etc.... En fait la multiplicité des vitesses se traduit par un élargissement des raies et la résolution initiale est augmentée de 2-3 ordres de grandeur grâce à l'étude simultanée d'un très grand nombre de fréquences. La Terre sur son orbite n'induit qu'une vitesse radiale de 0.1 m/s sur le Soleil. Mais des Terres plus proches induiraient un mouvement plus marqué. L'instrument le plus accompli à ce jour est le spectromètre HARPS (pour High Accuracy Radial velocity Planetary Search) installé au foyer d'un télescope de ∅3.6 m à La Silla, appartenant à l'ESO. HARPS obtient une précision record de 0.4 m.s^{-1}. La limite physique du phénomène dépend essentiellement de l'étoile. Pour les étoiles les plus calmes sans activité notable, cette limite se situe sous la limite instrumentale. On peut trouver une description sur le site de l'ESO (Pepe et al. 2002)

2.1.2 Astrométrie

L'astrométrie s'intéresse à la modulation de position d'une étoile sous l'effet d'une planète. C'est le même phénomène que pour la vélocimétrie radiale, mais ici c'est la position de l'étoile que l'on mesure.

Si dans le système précédent d est notre distance à l'étoile, la déviation angulaire de l'étoile vaut :

$$\sin\alpha = \frac{a_\star}{d} \simeq \alpha$$

En faisant intervenir les paramètres de la planète (Eq.2.1) :

$$\alpha = \frac{a}{d}\frac{m}{M+m}$$

Les masses et distance de l'étoile sont connues par ailleurs. On peut connaître a à l'aide de T par la 3ème loi de Kepler (Eq.2.2). On trouve alors la masse de la planète en fonction de l'écart angulaire. En pratique, la variation de position de l'étoile n'est pas mesurée de façon absolue, mais différentielle par rapport à un objet fixe, c'est-à-dire loin dans l'arrière-plan.

Cette méthode présente l'avantage d'être indépendante de l'angle sous lequel est vu le système. Mais bien qu'à masse égale l'éloignement d'une planète amplifie le déplacement de l'étoile, c'est un élément défavorable car la lenteur du mouvement le rend non-détectable. La Terre, vue à 10 parsecs (32 années lumières), ne déplace le Soleil que de 0.3μas (microseconde d'arc). Or les mesures depuis le sol sont limitées à une milliseconde d'arc. Elles devraient atteindre prochainement 10μas sur des champs réduits. Dans le domaine spatial, le projet SIM (Space Interferometric Mission) cible une précision de 4μas. L'ambitieux projet GAIA doit atteindre 1μas et mesurer un milliard d'étoiles de la galaxie. Il pourrait per-

mettre de détecter une centaine d'Uranus (jusqu'à $10m_{\oplus}$) dans les 10 pc qui nous entourent.

2.1.3 Méthode par réflexion

Le flux de lumière réfléchie par une planète varie avec sa surface éclairée (un "croissant") tout au long de son orbite. Même si l'étoile et la planète ne sont pas discernables ce croissant module le flux reçu. La variation de flux pour une planète de période P qui orbite autour d'une étoile de flux $L_\star$ est (Schneider 1999) :

$$L_p = \frac{AL_\star}{8}\left(\frac{r}{a}\right)^2 \Phi(t)$$

r est le rayon de la planète, A son pouvoir de réflexion isotrope (albédo). La phase Φ vaut :

$$\Phi(t) = 1 - \sin i \sin(2\pi t/T)$$

i est toujours l'inclinaison de l'orbite. L'amplitude de cette modulation décroît avec i. Mais dans la plupart des cas le taux de modulation est bien trop faible pour être détectable.

2.1.4 Observation directe

Imager directement une planète requiert un dispositif capable d'"éteindre" l'étoile de façon très efficace, coupant tout le flux lumineux pour une séparation angulaire de seulement 0.1 milli-arcseconde. Une occultation par un obturateur mécanique est très insuffisante et il faut avoir recours à l'optique en utilisant la l'interférométrie ou la coronographie. Ces principes sont rappelés par Schneider (1999). En voici quelques projets :

- La déformation contrôlée du miroirs d'un télescope pourvu d'un dispositif d'*optique adaptative* peut être utilisée pour faire interférer destructivement la lumière de l'étoile sur les différentes parties du miroir. Les rayons lumineux issus de la planète, suivant un trajet différent ne seraient pas atténués.
- Le *nulling* fait interférer deux télescopes situés à une distance l qui peut être ajustée pour obtenir une déphasage de 180°entre les télescopes. Cependant, issue d'une direction légèrement différente de celle de l'étoile, la lumière planétaire serait peu déphasée.
- Des projets d'interféromètres spatiaux sont en gestation tel DARWIN, une constellation de télescopes spatiaux volant en formation.

2.1.5 Lentilles gravitationnelles

Le phénomène de lentille gravitationnelle se produit lorsqu'un astre massif passe devant un objet observé en arrière-plan. L'astre courbe légèrement les

rayons lumineux émis par l'objet du fait de sa gravité et les concentre comme le ferait une lentille, provoquant un pic d'intensité lumineuse lors de son passage. Une planète en orbite autour de son étoile, bien que moins massive que cette dernière déviera également les rayons, faisant apparaître des pics secondaires. La Fig.2.2 montre la courbe de lumière de l'événement OGLE (pour Optical Gravitational Lensing Experiment Udalski et al. (1992)) 2003-BLG-235/MOA 2003-BLG-53 (Udalski et al. 1993)

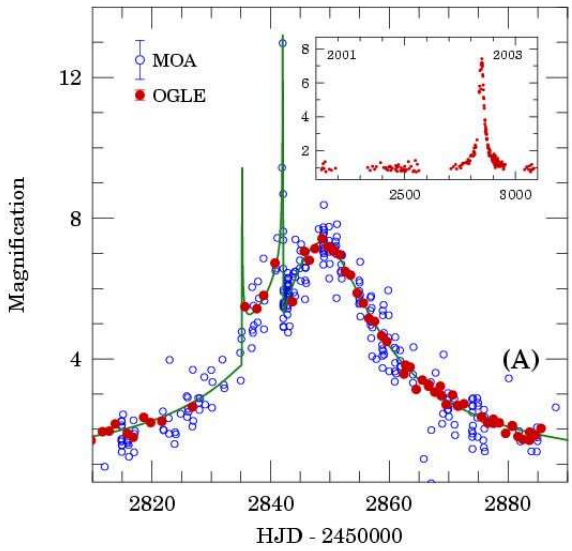


FIG. 2.2 – *Courbe de lumière de l'événement OGLE 2003-BLG-235/MOA 2003-BLG-53. La courbe théorique (ligne continue) explique bien les pics observés (cercles et disques)*

Ces phénomènes sont cependant très rares du fait de la précision d'alignement requise. Il est donc nécessaire d'observer un très grand nombre d'étoiles, par exemple en visant vers le bulbe de notre galaxie. En outre cette méthode ne permet qu'une étude statistique des planètes car les événements ne se produisent qu'une seule fois.

2.2 la Méthode des transits

Il s'agit de détecter des exo-éclipses par une mesure continue du flux lumineux d'une étoile. Une partie de ce flux est occultée par le passage d'une planète devant le disque de l'étoile (voir Fig.2.3), ce qui provoque une baisse temporaire et récurrente à la période orbitale. Le signal est plus accentué qu'avec la méthode

par réflexion car ici c'est directement le flux de l'étoile qui est affecté et non le flux réfléchi par la planète.

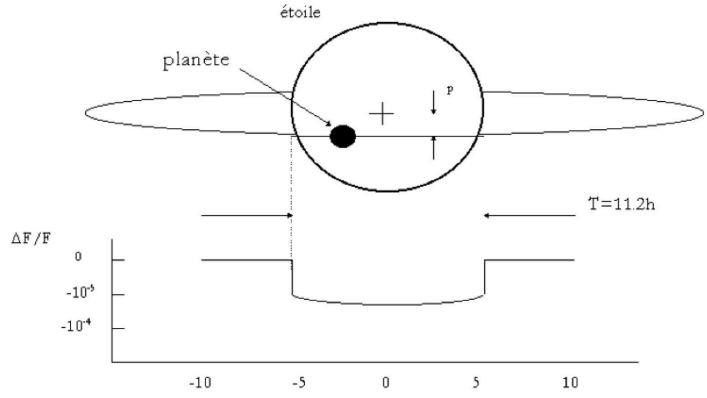


FIG. 2.3 – *Méthode des transits.*

Le transit d'une planète devant le disque de son étoile se traduit par une baisse du flux lumineux de l'étoile dépendante de la géométrique de l'événement. Pour une planète de rayon r passant devant une étoile de rayon R dont le disque est uniformément lumineux, l'amplitude du transit est égale au rapport des surfaces :

$$\frac{\Delta F}{F} = \left(\frac{r}{R}\right)^2 \tag{2.5}$$

Cette relation au premier ordre néglige la luminosité propre de la planète, hors d'atteinte des instruments. A $\left(\frac{r}{R}\right)$ donné, la présence d'un gradient de luminosité à la surface d'un disque stellaire jouera en faveur de la détection pour les planètes transitant près de son centre. La Terre produit un transit de 0.8×10^{-4} en passant devant le Soleil, Jupiter atteint 2.5×10^{-2}. Cet ordre de grandeur est accessible à la mesure. L'amplitude relative du transit a l'avantage d'être insensible à la distance observateur-étoile, mais nécessite un quasi-alignement de la ligne de visée avec le plan orbital de la planète.

2.2.1 Probabilité

Les orbites de rayon a qui peuvent produire une éclipse sont comprises dans la surface de la tranche de sphère, assimilée à un cylindre, de rayon a et hauteur $2R$ (voir Fig.2.4).

Le nombre d'orbites observables est donc proportionnel à $2\pi a \times 2R$. Leur fraction (ou probabilité) s'obtient en normalisant par la surface totale $4\pi a^2$. On obtient :

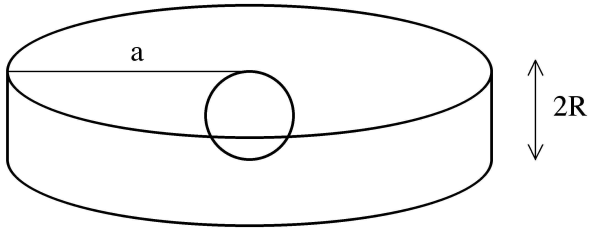


FIG. 2.4 – *Observabilité d'un transit.*

$$\mathcal{P} = \frac{R}{a} \tag{2.6}$$

En pratique $\mathcal{P} = 0.47\%$ pour la Terre et seulement 0.01% pour Jupiter. L'éloignement joue proportionnellement en défaveur des transits. Une orbite éloignée est également plus pénalisante du fait de l'allongement de la période de révolution. On augmente les chances de détection en suivant simultanément un grand nombre d'étoiles. Certaines expériences portant sur l'observation simultanée de toutes les étoiles d'un champ sont menées depuis le sol, telle OGLE ou WASP. D'autres le seront depuis l'espace comme la mission Européenne Corot ou la mission Américaine Kepler.

2.2.2 Durée du transit

Le paramètre d'impact p est la distance séparant la ligne de contact de la ligne équatoriale de l'étoile (voir Fig.2.3). L'éclipse parcourt la corde de longueur $l = 2R\sqrt{1-p^2}$ à la vitesse $v = 2\pi a/T$. Elle mettra donc un temps :

$$t = \frac{RT}{\pi a}\sqrt{1-p^2} \tag{2.7}$$

La table 2.1 résume les caractéristiques des transits Terrestre et Jovien.

TAB. 2.1 – *Caractéristiques-type des transits dans le système Solaire.*

Planète	$\Delta F/F$	$\mathcal{P}$	durée (h)	Période (année)
Terre	0.810^{-4}	0.47%	11.3	1
Mercure + Venus + Terre		2%		
Jupiter	2.510^{-2}	0.01%	29	11.9

2.2.3 Intérêt du spatial

Il faut donc une précision photométrique relative de quelques 10^{-4} sur 1h, temps caractéristique des transits, pour détecter le transit d'une planète tellurique. Depuis le sol, on ne peut guère espérer dépasser 10^{-3} en raison de la turbulence atmosphérique.

Idéalement il faudrait une observation continue pendant une très longue période de temps, ce que ne peuvent assurer les télescopes terrestres qui ne fonctionnent que la nuit. Certes on pourrait utiliser plusieurs télescopes répartis, mais la solution serait coûteuse et resterait soumise aux aléas météorologiques. Il faut en outre une stabilité des conditions de mesure équivalente à la précision recherchée, ce qui n'est pas le cas pour l'atmosphère.

A cause de la révolution de la Terre autour du Soleil, une région donnée vue depuis le sol ne reste visible durant plus d'une demi-nuit, que pendant trois mois. Le domaine spatial reste de prédilection pour la méthode des transits, car le Soleil peut rester pendant 6 mois situé à plus de 90 degrés d'une direction de visée.

La méthode des transits semble bien adaptée pour détecter de petites planètes, en particulier des planètes telluriques analogues à la Terre. Elle a été retenue dans le cadre de la mission spatiale Corot . Le présent travail s'inscrit dans ce contexte.

La limitation physique de la méthode provient essentiellement de deux causes : le bruit quantique ou photonique, et les variations de flux dues à l'activité propre des étoiles.

2.2.4 Le bruit photonique

Dans un flux lumineux faible de moyenne f, un photon peut atteindre le détecteur à n'importe quel moment. Le nombre de photons touchant le détecteur chaque seconde fluctue suivant une loi de Poisson de paramètre f. La probabilité de recevoir exactement i photons durant 1 seconde vaut :

$$P(i) = \frac{f^i.e^{-f}}{i!}$$

Selon une propriété connue, l'écart-type de la mesure est :

$$\sigma = \sqrt{f}$$

Le bruit photonique augmente en valeur absolue avec le flux, mais pas en valeur relative. En effet, le rapport signal à bruit (S/B) vaut :

$$r = f/\sqrt{f}$$

Il s'améliore comme $\sqrt{f}$. On a donc intérêt à collecter le plus de photons possible. Ceci s'obtient en préférant pointer les étoiles brillantes et en limitant les pertes à l'intérieur de l'instrument. Les bruits s'additionnant par leur variance, on convient généralement qu'aucun ne doit excéder le bruit photonique. C'est ce qui à été fait dans la mission Corot .

Les longueurs d'onde mesurées sont inférieures à $0.85\mu m$, bien moins que la dimension d'un pixel (13.5μm). On traite donc les photons comme des particules et non des ondes qui pourraient tomber dans deux pixels à la fois. Il n'y a donc aucun espoir de corriger le bruit photonique d'un pixel en cherchant sa contrepartie dans d'autres pixels.

2.2.5 La variabilité stellaire

Les étoiles sont le siège d'une activité intense se manifestant par des éjections de matière, de la granulation, des taches sombres etc.... Cette activité se signale par une variabilité du flux émis par l'étoile et qui perturbe le signal photonique. Les phénomènes en cause peuvent induire des changements sur des échelles de temps commensurables avec les transits. Par exemple l'apparition d'une tache sombre entraînée par la rotation de l'étoile peut imiter la signature d'un transit. La variabilité Solaire atteint 1/1000 et l'on attend 10 fois plus pour d'autres étoiles. Plus une étoile est chaude, plus elle est active. Pour les étoiles calmes, les temps caractéristiques attendus pour les phénomènes mis en cause sont supérieurs à ceux des transits et doivent pouvoir faire l'objet d'un filtrage numérique.

On peut estimer à partir de quelques hypothèses simples la proportion d'étoiles qui auront un niveau d'activité acceptable pour la détection des transits :

- le niveau d'activité diminue brutalement avec le nombre de Rossby $R_0 = P_{rot}/T_{conv}$, ou P_{rot} est la période de rotation de l'étoile et T_{conv} le temps de retournement de la convection. Une étoile en rotation rapide sera le siège de mouvements plus violents. Pour $R \leqslant 0.66$, la variabilité devient acceptable, comprise entre 0.1 et 0.001.
- La rotation d'une étoile diminue constamment avec son âge. Soderblom (1983) a vérifié par l'observation pour les Hyades que P_{rot} est proportionnel à la racine de l'âge de l'étoile. Le coefficient de proportion dépend du type spectral[1] [2].
- Pour un type spectral donné, on peut donc calculer l'âge à partir duquel $R \leqslant 0.66$. En comparant cet âge à la durée de vie de l'étoile, on obtient la

[1] Le type spectral est le classement des d'étoiles par gamme de températures. Il est désigné par une lettre O,B,F,G,K ou M par température décroissante

[2] La classe spectrale subdivision en dix chaque type spectral

fraction des étoiles calmes. On répète le calcul pour chaque type spectral présent dans le champ, correctement pondéré.

Les résultats sont présentés en table 2.2. On peut compter sur une proportion de 80% d'étoiles suffisamment calmes.

TAB. 2.2 – *Fraction d'étoiles ayant réduit leur activité, en fonction du type spectral*

Type spectral	$T_{\rm conv}$(jours)	$t_{\rm min}$(Gyr)	T(Gyr)	fraction
F8	3	0.2	4	0.95
G0	8	0.9	9	0.90
G2	10	0.7	1.	0.93
G5	20	2	1.2	0.83
G8	30	4	1.6	0.75
K0	30	3	3	0.9
K3	30	2.2		>0.9
K5	30	2		>0.9

Dans les étoiles les plus calmes dont fait partie notre Soleil, les principales variations photométriques attendues à l'échelle de temps des transits sont dues aux régions actives et aux mouvements turbulents. Des travaux sont menés pour mieux connaître ces phénomènes.

Modèle de rotation des régions actives

Les régions actives apparaissent et disparaissent sans cesse, entraînées par la rotation de l'étoile. Lanza et al. (2003) ont étudié la variation du flux Solaire dans les données acquises entre 1996 et 2001 par l'instrument VIRGO à bord de la sonde SoHO. Pour modéliser la variabilité, ils considèrent un nombre discret de régions actives réparties à la surface d'une sphère en rotation. En dehors de ces régions, la brillance du disque stellaire croît vers le centre sous l'effet de l'assombrissement centre-bord. C'est un effet d'incidence rasante qui fait qu'un rayon lumineux issu du bord du disque de l'étoile traverse une plus grande épaisseur d'atmosphère stellaire qu'un rayon issu du centre du disque. Les couches internes chaudes et intenses nous sont donc masquées. Un exemple de paramétrage de leur modèle est donné figure 2.5.

Lanza et al. (2003) ont utilisé trois régions actives renouvelées tous les sept jours en taille et position pour tenir compte de leur évolution propre. A l'image des taches Solaires, les régions sont formées d'une zone sombre de moindre convection cerclée de facules brillantes. Leur simulation comporte onze paramètres libres : 6 de position, 3 d'aire des surfaces actives, 1 de brillance

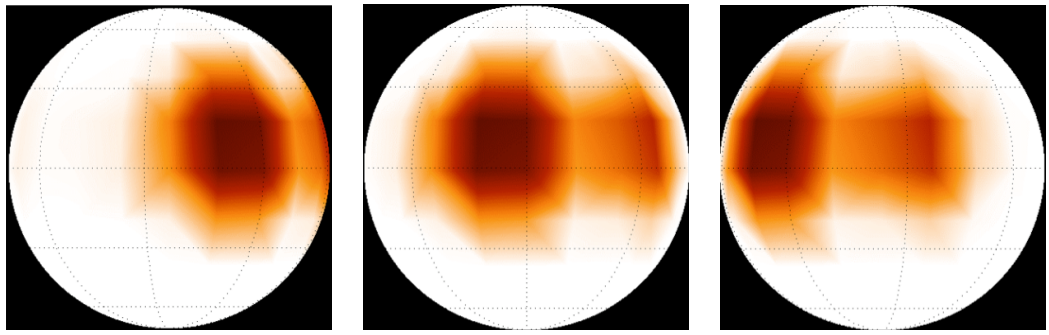


FIG. 2.5 – *Déplacement de taches sombres sur une sphère non homogène en rotation.*

du disque d'arrière-plan, le dernier est la période de rotation Solaire. Ils ont ajusté au mieux ces paramètres en minimisant une statistique du χ^2. L'accord a permis de reproduire de manière fiable la variabilité bolométrique du Soleil sur une échelle de sept à dix jours, en dehors de son maximum d'activité. L'erreur résiduelle est de l'ordre de 2.10^{-4}. La figure 2.6 détaille cet ajustement pour deux périodes de quelques semaines. En retranchant le modèle des données réelles, cette technique réduit le bruit d'activité Solaire d'un ordre de grandeur.

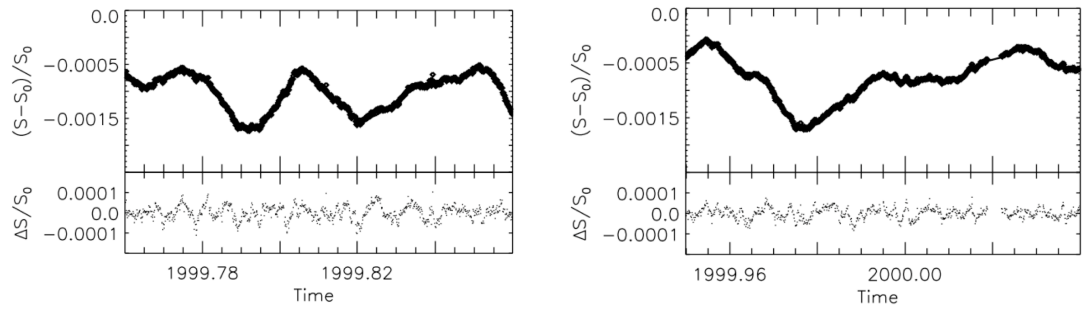


FIG. 2.6 – *Accord entre simulation et flux Solaire pour deux plages de temps. La variation relative de luminosité figure dans le cadran supérieur. Les points de mesure sont en gras, la simulation est la ligne continue qui émerge à* $t = 2000.02$*. Les points du modèle ne sortent pratiquement pas des points mesurés. Le cadran inférieur contient le résidu.*

Outre sa capacité à réduire l'effet de la variabilité stellaire dans les courbes de lumière, cette technique devrait fournir des indications sur la rotation propre de l'étoile ainsi que l'inclinaison de son axe. Pour l'utiliser dans la recherche de transits, il faut rester vigilants à ce que l'ajustement du modèle n'intègre pas le transit lui-même comme une configuration particulière de l'activité.

Modèle de spectre de granulation

L'autre principale source de variabilité stellaire est la convection. Les différentes échelles des cellules se traduisent par différents temps caractéristiques (voir Tab.2.3).

TAB. 2.3 – *Phénomènes convectifs types.*

Activité	T(s)	f	
Régions actives	1 à 3.10^5	5	μHz
super granulation	3 à 7.10^4	20	μHz
méso granulation	8 000	125	μHz
granulation	200 à 500	3	mHz
points brillants	70		

Les différentes échelles de granulation contribuent indépendamment au spectre de l'étoile. Harvey et al. (1993) considèrent que chaque granulation apporte au spectre d'irradiance la contribution :

$$P(\nu) = \frac{A}{1 + (2\pi\nu T)^b}$$

qui correspond à la décroissance exponentielle d'une fonction d'autocorrélation de temps caractéristique T. A est l'amplitude et b un coefficient à déterminer.

Pour étudier l'influence de la granulation sur la détection des transits, Aigrain et al. (2003) ont élaboré une méthode qui reproduit efficacement le spectre des mesures VIRGO SoHO sur la période de 1996 à 2001. L'objectif est de simuler des courbes de lumière pour tester la robustesse des algorithmes de détection. Dans leur méthode, ils considèrent le spectre d'une section de courbe de lumière SoHO de longueur $L = 180$ jours. Ils l'ajustent avec un premier jeu de paramètres A_1, T_1, b_1, puis ils affinent l'accord à l'aide du jeu A_2, T_2, b_2 et ainsi de suite jusqu'à ce que l'apport d'une nouvelle composante n'apporte pas d'amélioration. Ils appliquent le même traitement pour la section suivante, obtenue en décalant l'origine de 20 jours et en utilisant les A_i, T_i, b_i comme nouvelles valeurs initiales. La figure reproduite ici (Fig.2.7) montre un exemple d'ajustement de spectre. Il s'avère que trois composantes sont suffisantes pour modéliser le spectre Solaire.

Mené dans plusieurs bandes de couleur, l'exercice souligne que pour le Soleil l'effet de la granulation est plus marqué dans le vert et le bleu que dans le rouge.

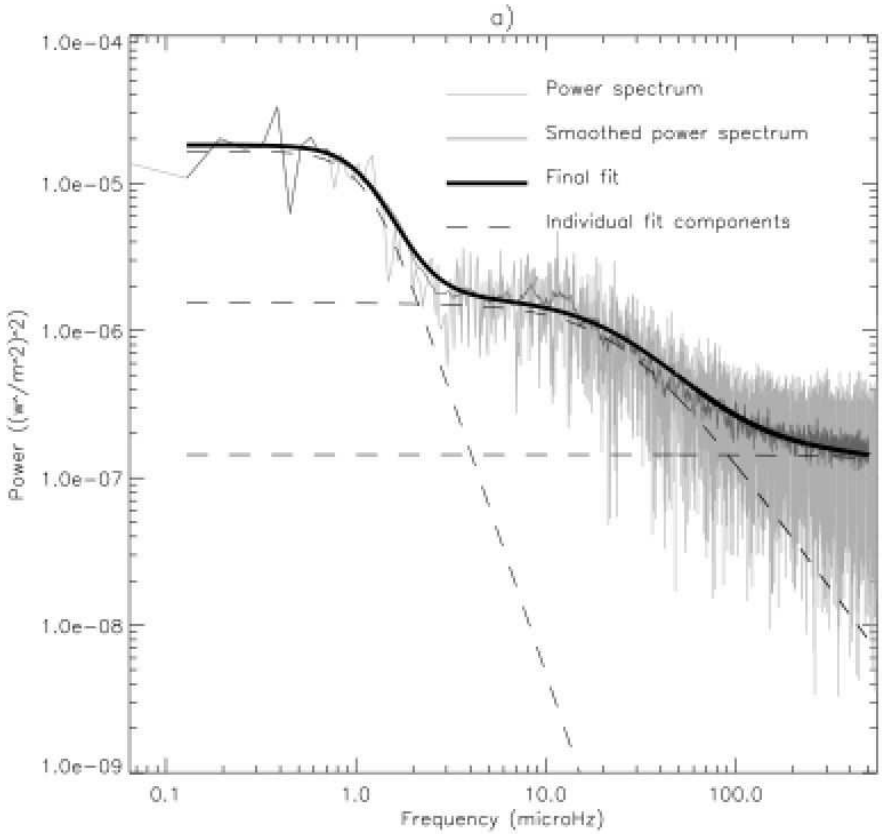


FIG. 2.7 – *Exemple d'ajustement du spectre Solaire par 3 composantes.*

Chapitre 3

La recherche de planètes extrasolaires avec Corot

3.1 Présentation de la mission

Corot (pour COnvection ROtation and planetary Transits) est une mission spatiale pour la détection de planètes par la méthode des transits. Elle observera 60 000 étoiles de manière continue pendant 5 périodes de 150 jours. Elle offre une précision photométrique atteignant 7.10^{-4} pour des étoiles de magnitude 12 à 15.5 en flux intégré pendant une heure.

Corot fait partie des 'mini missions' du CNES, qui en assure la maîtrise d'ouvrage et le lancement. C'est un projet Européen qui a vu le jour sous l'impulsion du Laboratoire d'Etudes Spatiales et d'Instrumentation en Astrophysique (LESIA, à Meudon), de l'Institut d'Astrophysique Spatiale (IAS, à Orsay) et du Laboratoire d'Astrophysique de Marseille (LAM), cadre de la présente thèse. Corot embarque deux expériences de photométrie de haute précision : L'une de sismologie stellaire, consacrée à l'étude des couches profondes par la mesure des modes d'oscillation de l'étoile (le 'chant' des étoiles) sous l'effet de la pression et de la gravité (Baglin et al. 2001) ; l'autre de détection d'exoplanètes par la méthode des transits [1]. Les observations sont réparties à raison de 5 champs de 12 000 étoiles pendant les 2ans $\frac{1}{2}$ que dure la mission

L'instrument Corot est un petit télescope de ∅27cm. Le lancement est prévu à Baïkonour au $2^{ème}$ semestre 2006 par une fusée SOYUZ. L'orbite polaire inertielle permet de conserver une direction d'observation fixe pendant six mois sans être ébloui par le Soleil ni masqué par la Terre (voir Fig. 3.1). L'orbite située

[1] Les sites officiels de Corot se trouvent sur `http ://corot.oamp.fr/` et `http ://smsc.cnes.fr/COROT/`. Le site consacré aux exoplanètes peut être consulté sur `http ://media4.obspm.fr/exoplanetes`

à 896 Km d'altitude est parcourue en 1h42min. Un observateur situé sur Corot verrait l'horizon terrestre tourner autour de la ligne de visée en 1h42, à 20 degrés d'elle, tout en défilant continûment. Ce défilement ferait alterner les périodes jour et nuit une fois par orbite. Le Soleil se déplacerait lentement à l'arrière passant en six mois d'un côté à l'autre. Après six mois, le Soleil commence à passer à l'avant. Corot est pivoté de 180° et un nouveau semestre d'observation débute. Ni la lune ni les planètes du système solaire ne passent près de cette ligne de visée.

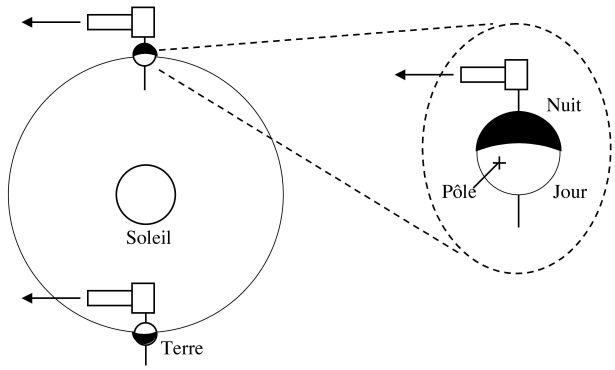


FIG. 3.1 – *Corot pointe perpendiculairement au plan de son orbite. La Terre ne passe jamais dans le champ. Le Soleil reste dans le "dos" de Corot durant un semestre. Le satellite est pivoté de 180après chaque phase d'observation.*

L'altitude du satellite est un compromis. Trop basse l'horizon lumineux serait près de l'axe de visée, trop haute le champ magnétique terrestre ne protégerait plus Corot contre les rayons cosmiques ionisants, principalement rencontrés lors la traversée de l'anomalie de l'atlantique sud (SAA).

Le satellite est constitué de l'instrument Corot porté par une plate-forme générique PROTEUS (Plate-forme Reconfigurable pour l'Observation, pour les Télécommunications et les Usages Scientifiques) développée par le CNES. C'est elle qui assure les fonctions de communication, pointage, énergie etc.... Les panneaux solaires pivotent tous les 15 jours pour suivre le mouvement apparent du Soleil.

L'orientation du satellite dans les trois axes est maintenue constante par un mécanisme dynamique. Toute variation est détectée par les senseurs stellaires et gyromètres de la plate-forme. Celle-ci déclenche une action contraire en modifiant la vitesse de roues à inertie dans les trois axes. Le couple moyen des forces externes n'étant pas nul dans le repère du satellite, la vitesse des roues à tendance à augmenter en permanence. Pour les ralentir, on les couple périodiquement avec le champ magnétique terrestre à l'aide de magnéto-coupleurs.

Cette opération s'appelle la désaturation. La constance du pointage est un point critique pour Corot car les fluctuations de visée induisent un bruit de mesure : le bruit de jitter (détaillé au chapitre 5). PROTEUS offre une précision standard de 1 à 2 secondes d'arc (notées 2"), ce qui correspond à un "flottement" supérieur demi-pixel, excessif pour Corot . Une des spécificité de cette mission est d'améliorer ce pointage en confiant la mesure de la consigne au télescope lui-même. On espère ainsi un jitter inférieur a 0.15" d'arc, soit $1/10^e$ de pixel.

Moins de 6% du temps sera perdu pour la mesure, principalement à cause de la rotation des panneaux solaires, de la traversée de la SAA, et des calibrations.

La table 3.1 indique les principales caractéristiques de la mission.

TAB. 3.1 – *Caractéristiques techniques de la mission Corot .*

Masse	entre 570 et 630 kg
Masse Charge Utile	$\simeq$ 270 kg
Hauteur	4100 mm
Diamètre	1984 mm
Puissance électrique	380 W
Précision du pointage	0.15 arcsec
Télémétrie	900 Mbit/jour
Orbite	896 Km
Période	1h 42min
Durée d'exposition	512s (8.5 min), en sommant 16 poses de 32s
Nombre de cibles	60 000, à raison de 5 champs de 12 000
Durée d'observation continue	6 mois par champ
Durée totale de la mission	2 ans 1/2
$\varnothing$ pupille d'entrée	27cm
champ exoplanètes	3.5 deg^2
m_V cibles	de 10.5 à 15.5
Longueurs d'onde	370 à 950 nm
Sensibilité	$\Delta F/F = 7.10^{-4}$ à $m_V = 15.5$
Réjection du baffle	10^{-13}
Objectif	6 lentilles dioptriques
CCD	2048 $\times$ 2048
Taille pixel	13.5μm
Jitter	0.5 arcsec (0.2 pixel)

3.2 l'Instrument

L'instrument Corot est schématisé figure 3.2. Le télescope forme une image sur 4 capteurs CCD indépendants de 2048×2048 pixels de $13.5\mu m$ dont deux sont dédiés aux exoplanètes. Le CCD est éclairé par l'arrière et fonctionne en mode transfert de trames. Ce capteur est stable et reste linéaire sur une grande échelle. La même conformation permet de mesurer des phénomènes sur plus de deux ordres de grandeur. Suivant les étoiles, le capteur reçoit un flux intégré de quelques centaines à près d'une dizaine de milliers de photons par seconde (voir table 3.2). Ces photons sont accumulés durant 32s sous forme d'électrons, puis la charge électrique résultante est acheminée vers une électronique de mesure pour y être numérisée. Les mesures sont sommées sur 512 s ($\simeq$ 8.5 min), temps de pose élémentaire, puis transmises au sol lors du passage suivant à portée d'une antenne. Le bruit d'un CCD diminue avec sa température. Celle-ci est régulée vers -40℃ au 1/100° près. L'électronique de mesure est maintenue à +20℃.

TAB. 3.2 – *Nombre de photons (ou photoélectrons notés e^-/s) capturés en fonction de l'étoile.*

Teff mv	4500 K	5750 K	6500 K	8000 K
10.5	135 667	105 268	98 358	88 811
12	34 078	26 442	24 706	22 308
14	5 427	4 165	3 912	3 881
16	860	660	620	615

Les champs visés par Corot s'étendent sur 3.5 deg^2, soit 8 fois la surface de la pleine lune. Ils sont choisis dans 2 directions opposées de 180 ° : Le centre de la galaxie et l'anticentre. Ces champs sont suffisamment denses pour contenir 12 000 cibles brillantes de magnitudes comprises entre 12 et 16[2].

3.2.1 Le baffle

L'altitude n'est que $2/10^e$ du rayon de la Terre ; l'horizon terrestre se trouve donc situé à 20° seulement de la ligne de visée (Fig. 3.3). A la précision requise, la lumière parasite réfléchie et diffractée sur les organes mécaniques du satellite devient prépondérante. Un baffle d'entrée est chargé d'éliminer toute lumière hors d'axe. C'est un cylindre absorbant de forme allongée situé dans le prolongement

[2]La magnitude visuelle d'une étoile dont on reçoit le flux f est $m = -2.5 \log \frac{f}{f_0}$, f_0 étant un flux de référence. Le flux est divisé par 10 quand la magnitude augmente de 2.5 unités.

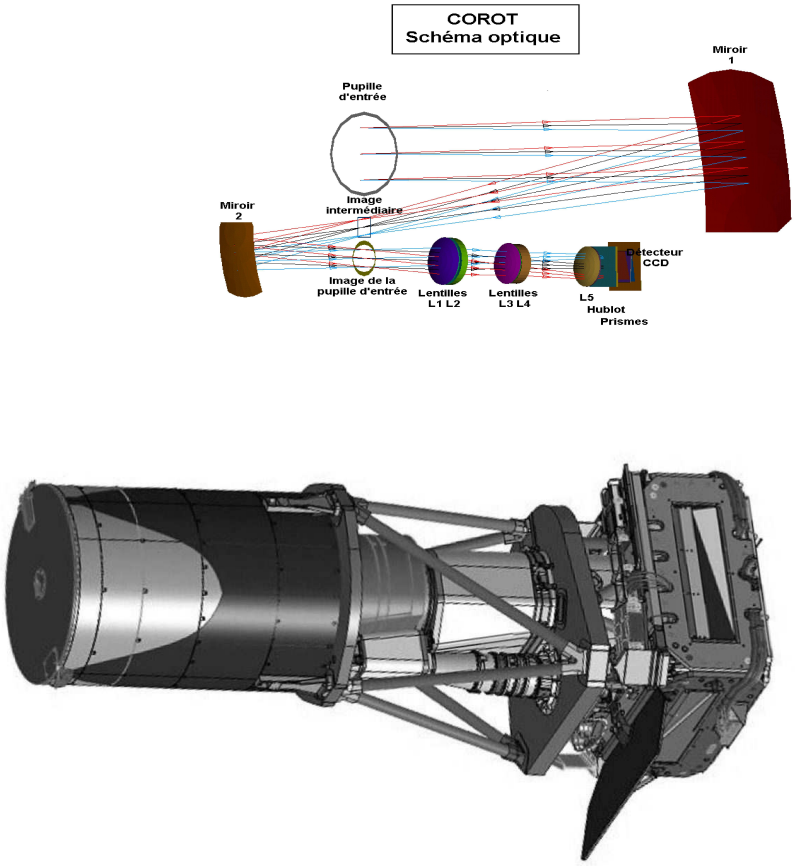


FIG. 3.2 – *Chemin optique (en haut) et Instrument Corot (en bas). La lumière entre par la gauche. Elle est filtrée par le baffle (cylindre de grand diamètre) après que l'obturateur (couvercle) à été ouvert. Le miroir primaire (extrémité droite) la réfléchit sur le miroir secondaire (au pied du baffle). Elle pénètre alors dans l'objectif dioptrique (tube oblique inférieur entre les armatures) contenant les six lentilles qui la guident vers le bloc focal. Celui-ci, de la taille d'une boîte à chaussures est situé dans le plateau à droite des armatures. Il contient le prisme et les CCDs, protégés des rayons cosmiques par un blindage. La case à équipements (bloc de droite) contient les boîtiers électroniques d'acquisition, de régulation thermique au 1/100°C, et l'électronique de traitement des données. Le plan inférieur saillant sous la case est le radiateur du bloc focal*

du télescope. Tout rayon lumineux incident indésirable y subit au moins trois réflexions absorbantes sur les chicanes internes. L'atténuation record atteint 10^{-13} pour les photons terrestres.

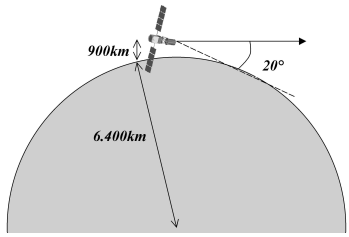


FIG. 3.3 – *Horizon de Corot*

Un obturateur à l'entrée du baffle protège le détecteur de la vue directe du Soleil lors des manoeuvres de mise en orbite du satellite. Il permet aussi de faire les étalonnages de début de mission. Après stabilisation de Corot en attitude, l'obturateur s'ouvre définitivement.

3.2.2 La défocalisation de l'image

La luminosité des étoiles d'une région dense s'étend sur plusieurs ordres de grandeur. Pour une étoile ponctuelle trop brillante, l'excès de lumière dépasse la capacité de 120 000 photoélectrons que peut contenir un pixel : C'est la saturation. L'étoile est non seulement perdue pour la mesure, mais le trop plein d'électrons compromet les étoiles voisines en s'épanchant de part et d'autre dans la colonne. Pour limiter cet effet, la tache image est volontairement étalée par insertion d'un système disperseur. Ainsi les photons d'une étoile recouvrent en moyenne 80 pixels. Corot atteindra une précision photométrique inégalée, mais sera volontairement myope.

La défocalisation à d'autres avantages, elle protège contre :
- La différence de sensibilité entre pixels, nommée PRNU (Pixels Response Non Uniformity),
- Les différences au sein d'un même pixel,
- La perte des étoiles trop proches de l'interstice inter-pixels.

La défocalisation repousse la magnitude saturante à $m_V \lesssim 12$, suivant le spectre de l'étoile, sa position et son entourage. En approchant de cette limite, le

CCD perd graduellement sa linéarité.

3.2.3 Le système disperseur

Un bi-prisme est inséré sur le trajet lumineux. Il disperse spatialement les couleurs afin de distinguer la signature d'un transit planétaire des fluctuations intrinsèques l'étoile. En effet, un transit est un événement essentiellement géométrique dont la signature lumineuse est similaire dans les trois bandes de couleurs rouge, vert, bleue que délivre Corot . A l'inverse une fluctuation stellaire, telle une tache sombre, a une origine thermique qui se traduit par une variation chromatique.

3.2.4 Les PSFs de référence

Une étoile Corot a une extension angulaire de l'ordre du nano-arcsec, un million de fois inférieure au pixel. C'est donc un point source optique, superposition ponctuelle des longueurs d'onde du spectre stellaire. Sa tâche image I est la réponse impulsionnelle F de l'instrument, nommée PSF (de l'Anglais Point Spread Function) sommée sur le spectre :

$$I(x,y) = \int_{\lambda} a(\lambda)\delta(x,y) * F_{\lambda}\, d\lambda$$

ou δ désigne le Dirac positionné sur l'étoile, $a(\lambda)$ le spectre en longueurs d'onde. La répartition des photons dans la PSF n'est pas uniforme, les composantes rouges (grands λ) sont localisées à droite de la PSF (CCD vu par l'arrière) et contiennent la plupart de l'énergie. La composante verte est au centre et la bleue à gauche. La position sur le CCD modifie I.

La figure 3.4 résume la transformation subie par la lumière le long du chemin optique de Corot .

Les taches images du champ exoplanètes comptent 50 à 120 pixels émergeant du bruit, selon la magnitude de l'étoile.

Les deux méthodes employées pour connaître la PSF de Corot sont la modélisation et la mesure expérimentale. Il est difficile de connaître les PSFs de vol au niveau de précision exigé. Avant le tir, on travaille sur des PSFs simulées et complétées de quelques vérifications expérimentales. On utilise le logiciel de simulation ZEMAX qui envoie des rayons virtuels à travers les éléments optiques du modèle de l'instrument, miroirs, lentilles etc..., et calcule leur impact dans le plan du CCD. Le logiciel explore l'espace des paramètres de position dans la

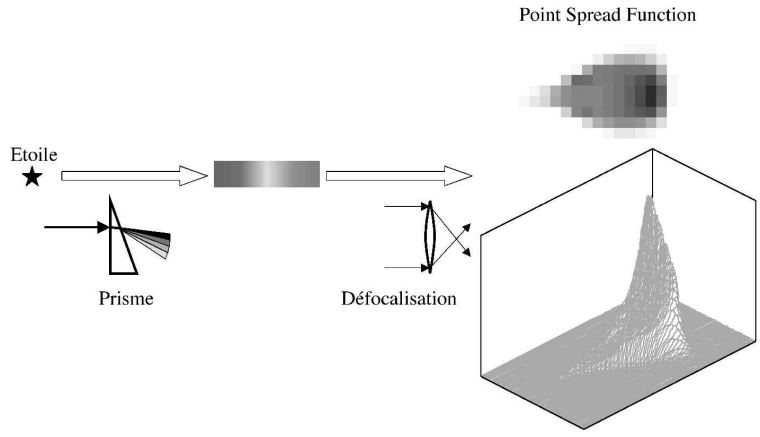


FIG. 3.4 – *Schéma du chemin optique. Après étalement du spectre par le bi-prisme, les rayons monochromatiques atteignent le CCD avec une défocalisation qui dépend de leur longueur d'onde. La tache image d'une étoile n'est pas uniforme, l'énergie incidente est principalement contenue dans la composante rouge, sur la droite.*

pupille, d'angle et de longueur d'onde du rayon.

L'échantillonnage en longueur d'onde couvre la gamme $350 - 1050$ nm par pas de 1 nm, chaque PSF étant calculée pour 18 positions réparties sur les deux CCDs. Afin de réduire le nombre d'échantillons nécessaires de 700 à ~ 170, Llebaria et al. (2004) ont utilisé un pas non constant plus espacé vers les grandes longueurs d'onde où la PSF varie lentement. Les longueurs d'onde manquantes sont aisément retrouvées par interpolation 1D des pixels. Ces longueurs d'onde sont combinées en 16 spectres stellaires de référence dont la composition en longueur d'onde est indiquée par Pickles (1998). Les types spectraux correspondants sont indexés uniquement par T, température de la photosphère. La PSF est une fonction continue, mais elle est délivrée sous forme d'une imagette sur-échantillonnée au pas de $1/5^{\text{e}}$ de pixel Corot .

La mesure expérimentale des PSFs est prévue en début de mission. Corot transmettra au sol quelques clichés complets acquis dans les conditions d'un bon S/B . Ces images de vol permettront de remonter aux vraies PSFs en se basant sur les PSFs simulées. La meilleure information proviendra des étoiles brillantes et isolées. L'image de vol à l'avantage d'être la vraie mesure in situ, mais son usage présente néanmoins deux difficultés :

– Après compensation du traînage (voir le paragraphe sur les bruits, §5.2.6), il faut tout d'abord séparer la PSF d'une étoile de celle du fond. Ce fond contient les étoiles d'arrière-plan, la composante continue et les bruits. On

commence classiquement par affiner la PSF en utilisant les étoiles les plus brillantes. Les étoiles d'arrière-plan sont ajustées à partir du résidu de la soustraction. Le fond continu est ajusté par un polynôme de faible degré traduisant des variations lentes.

– L'autre difficulté vient de la pixellisation. Le problème consiste à partir d'une image discrète à déduire la PSF continue sous-jacente. Cette déduction est nécessaire car la même PSF sert à plusieurs étoiles distantes d'un nombre de pixels non entier.

3.2.5 Position de référence

Pour positionner une PSF polychromatique sur le CCD, il faut établir une correspondance avec un de ses points pris comme référence. Avec les télescopes terrestres soumis à la turbulence atmosphérique, l'étoile est au centre de sa PSF concentrique, maximum d'une Gaussienne 2D centrée sur l'étoile. Par extension si la PSF n'est pas symétrique, on choisit de continuer à centrer sur le maximum. Mais dans Corot la position de la PSF monochromatique dépend de la longueur d'onde à cause du bi-prisme. Or les coordonnées du point de référence de la PSF sur le CCD ne doivent pas dépendre du spectre de l'étoile. On choisit le point maximum de la PSF monochromatique à $\lambda = 650nm$, longueur d'onde à laquelle la déviation par le prisme est nulle.

3.3 Photométrie

Les deux principales méthodes en la matière sont la photométrie par ajustement de PSF et la photométrie d'ouverture. C'est cette dernière qui est utilisée dans Corot .

3.3.1 Photométrie par ajustement de PSF

Les grands principes en sont rappelés par Debray (1982). En présence de bruit, l'image d'un objet O à travers un instrument de réponse F vaut

$$I = O * F + N$$

N représente le bruit d'origine photonique, thermique, électronique ou environnemental. Le bruit prend la forme d'un fond brillant et irrégulier qui varie entre deux expositions. Dans l'espace de Fourier, cette équation s'inverse en

$$\tilde{O} = \frac{\tilde{I} - \tilde{N}}{\tilde{F}}$$

On ne peut pas déduire O de cette manière, car pour les fréquences spatiales où $\tilde{F}$ est faible le terme $\frac{\tilde{N}}{\tilde{F}}$ créera une erreur importante.

Dans le cas d'une optique achromatique et linéaire, l'image d'un champ stellaire est en première approximation la répétition d'une même PSF I affectée de coefficients affines différents. Ces coefficients constituent la mesure de flux recherchée.

Dans un champ dense les PSFs se chevauchent, le recouvrement pouvant même être total. Un point donné de l'image R est la somme des contribution des k PSFs voisines. Supposons que k PSFs contiennent le point (x, y)

$$R(x, y) = N(x, y) + A.I + \sum_{i=1}^{k} A_i I_i(x - x_i, y - y_i)$$

où A est l'amplitude de l'étoile cible, I sa PSF et A_i l'amplitude de la PSF I_i centrée en (x_i, y_i) des étoiles de fond. Les I étant connues, la résolution du problème consiste à retrouver le meilleur A par des méthodes statistiques. Cet ajustement passe par l'évaluation des A_i. Il faut tenir compte du fait que les I_i, x_i et y_i sont connus avec une précision limitée.

Cette approche ne convient pas à Corot pour plusieurs raisons :

- Le débit de transmission des mesures (télémesure) est insuffisant. Il faudrait 5Gbit/jour pour transmettre 12 000 imagettes 15×10 toutes les 8.5 minutes alors qu'on ne dispose que de 900Mbit/jour. On ne peut comprimer ces données à cause de la puissance de calcul limitée disponible à bord. De plus des données comprimées sont fragiles car l'altération d'un seul bit peut entraîner la perte d'un bloc, d'autant plus grand que le flot est plus comprimé. L'ajout de codes correcteurs augmenterait l'encombrement de la bande passante et le travail du microprocesseur.

- Pour être efficace, cette photométrie requiert une connaissance précise des PSFs. Les PSFs de vol seront différentes des PSFs simulées à cause des changements de conditions, ou des déformations subies lors du tir. Or, dans Corot , les PSFs sont difficiles à mesurer in situ avec précision car elles sont très dépendantes du type spectral, de la position et du centrage par rapport aux pixels. Par conséquent une PSF ne se reproduit pas deux fois dans les mêmes conditions, on ne peut donc pas l'affiner en moyennant plusieurs réalisations. Le type spectral lui-même n'est connu qu'avec une précision limitée. En effet il est mesuré à travers l'atmosphère par des télescopes terrestres et la mesure ne porte que sur 4 bandes de couleur

standard B,V,r et i (voir table 3.3).

En outre pour avoir un bon résultat, il faut résoudre non seulement les PSFs des étoiles cibles, mais également celles des étoiles de fond qui y participent.

Pour terminer, les PSFs varient dans le temps au cours de l'orbite à cause des déformations dues aux différences entre la partie diurne et nocturne de l'orbite (la "respiration"), et à la gigue de pointage (le "jitter").

3.3.2 Photométrie d'ouverture

On définit tout d'abord une ouverture autour de la PSF : le *masque photométrique*. La photométrie d'ouverture consiste à sommer le flux dans les pixels du masque et à en retrancher le fond estimé. Le flux recherché est donné par :

$$f = \sum_{i=1}^{N} f_i - N \times b$$

où les f_i sont les flux dans les N pixels que compte le masque. b est la valeur commune du fond estimé. Cette photométrie est très simple. Elle comporte cependant quelques points délicats :

- Le résultat dépend de la forme du masque. Dans les systèmes classiques, le masque est soit un disque centré (pour la turbulence atmosphérique), soit une ligne isophote. Pour Corot , la détermination des masques est compliquée par d'autres facteurs. Un masque inadapté bruitera la mesure. La détermination des masques est l'objet de la section 5.7.2.

- L'estimation du fond peut être faite par extrapolation du profil d'intensité sur des couches concentriques autour de l'étoile, ou par l'histogramme des pixels environnants. Mais le problème devient délicat quand le fond contient à son tour des étoiles, comme c'est le cas pour les champs denses. Ces étoiles vont être incorporées au fond ou à l'étoile. Heureusement pour Corot , la mesure ne porte que sur des variations temporelles. Les sources secondaires et le fond s'en trouvent automatiquement soustraits. En revanche leur variabilité dans le temps persistera et pourra être une source d'ambiguïtés.

Dans la photométrie d'ouverture pratiquée avec Corot , la sommation est effectuée en vol et le fond est retranché par post-traitements. Ce dernier est

estimé à partir de fenêtres noires ne contenant pas d'étoiles. Le masque photométrique est virtuel, obtenu par programmation de l'électronique de lecture. Cette technique réduit considérablement le volume des données, chaque étoile se résume à 24 bits/8.5 min. Elle est également adaptée au bruit de photons dominant, qui se moyenne dans la sommation.

Afin de distinguer les couleurs, le masque des 5 000 cibles les plus brillantes est subdivisé en trois sous-masques dits colorés, bleu, vert, rouge sommés indépendamment. Ces définition de couleurs ne correspond pas aux bandes standard (voir Tab. 3.3), il s'agit d'une définition propre à Corot , basée sur une répartition de l'énergie d'environ 40%, 30% et 30% variable d'étoile à étoile.

Une variante de la photométrie d'ouverture est proposée pour le futur projet Kepler, où les pixels sont pondérés individuellement avant sommation à bord. Ceci permet par exemple d'affaiblir ponctuellement une étoile d'arrière plan. On trouve une description de ce projet par Duren et Al (Duren et al. 2004) ou sur internet [3].

La détermination des masques obéit à de nombreuses contraintes et fait l'objet de la première partie de cette thèse.

3.3.3 Sélection des cibles

Toute étoile n'est pas forcément une bonne cible. Tout d'abord il faut que la région pointée satisfasse simultanément les exigences du $2^{ème}$ objectif scientifique de Corot qui s'intéresse à l'astéro-sismologie. Celle-ci exige qu'au moins une étoile très brillante ($m_V \leqslant 6$) isolée soit positionnée sur l'un des deux CCDs dédiés. Pour la voie exoplanètes qui utilise les deux autres CCD nous avons vu précédemment que les meilleurs conditions de type spectral et de magnitude étaient obtenues pour des étoiles naines F,G,K,M avec $m_V \leqslant 15.5$. On peut aussi ajouter un critère de contamination (le taux de photons dans la PSF provenant d'étoiles de fond ou voisines) inférieur à 10%, imposé par la densité importante des champs du centre et de l'anticentre galactique.

Plusieurs régions peuvent réunir ces conditions. Pour y déterminer précisément le meilleur pointage, un recensement systématique à été mené à partir du sol avec la caméra grand champ du télescope Isaac Newton situé à La Palma[4]. Des millions d'étoiles ont été mesurées jusqu'à la magnitude $m = 20$ dans 4 bandes de couleur standard B, V, r et i (voir table 3.3), permettant par différence d'accéder au type spectral et d'avoir une idée de la classe de luminosité. Bien

[3] `http ://www.kepler.arc.nasa.gov/`

[4] Informations techniques `http ://www.ing.iac.es/Astronomy/telescopes/int/index.html`

que les étoiles faibles ne soient pas des cibles potentielles, il est important de les connaître car ce sont les plus nombreuses et leur lumière participe au fond de ciel. Un risque important lié à ce fond stellaire est la présence d'éclipses d'étoiles binaires. Ces éclipses peuvent, en effet, mimer un transit sur l'étoile principale et engendrer des ambiguïtés.

TAB. 3.3 – *Photométrie standard.*

domaine spectral	indice de couleur	$\lambda(\mu m)$	$\Delta\lambda(\mu m)$
UV	U	0.36	0.068
bleu	B	0.44	0.098
visible	V	0.55	0.089
rouge	R	0.70	0.22
proche IR	I	0.90	0.24

Les millions d'étoiles collectées sont stockés dans une base de données (EXODAT) qui est développée au LAM et qui permet d'établir les profils précis de chaque position. La densité d'étoiles dans le champ étant très inhomogène, on cherche à affiner le compromis entre grand nombre de cibles et encombrement excessif. Un outil de sélection des cibles dédié aux deux objectifs scientifiques de Corot à été développé à cette fin, il s'agit du simulateur COROTSKY (voir sites internet). Ce logiciel permet de visualiser les CCDs sur le ciel, et de faire toutes sortes de statistiques. Pour un pointage donné, COROTSKY fournit la liste et les caractéristiques des étoiles retenues comme cibles à destination du choix des masques.

3.4 Capacité de détection

La détectivité de Corot en termes de types de planète dépend de nombreux paramètres planétaires, dont certains sont mal connus. On rappelle le rôle que jouent les principaux d'entre eux :

- Le rayon de la planète est de première importance. Les plus grandes sont plus facilement détectables mais probablement gazeuses. On leur préfère les petites planètes potentiellement tellurique mais moins faciles à détecter.

- La magnitude visuelle de l'étoile. Les étoiles plus brillantes offrent un meilleur S/B , donnant accès à des planètes plus petites. Mais les étoiles moins brillantes ont l'avantage du nombre, qui double à chaque magnitude

supplémentaire (voir Fig. 3.5).

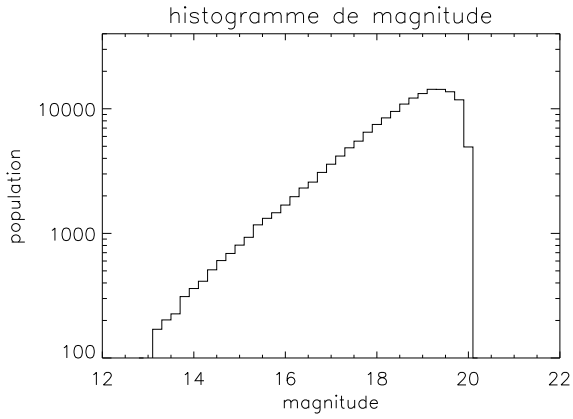


FIG. 3.5 – *Répartition des magnitudes en échelle logarithmique dans un champ Corot de l'anticentre galactique, jusqu'à la magnitude de coupure du catalogue. Le nombre d'étoiles progresse d'un facteur 2.2 à chaque pas de magnitude*

– La taille de l'étoile. Un transit sera plus marqué sur une étoile de petit diamètre. En revanche un plus grand éclat dû à un accroissement de la surface émissive n'est pas un avantage. Le flux supplémentaire n'est pas affecté par le transit et ne fait qu'augmenter le bruit de photons. Les étoiles naines de la séquence principale (le Soleil est considéré comme tel) présentent en outre l'intérêt d'une activité moindre.

– La distance de la planète. Une planète trop éloignée aura des transits moins fréquent et moins de chances d'être vue par la tranche (voir Eq. 2.6). Trop proche elle sera plus chaude car soumise à un flux intense.

– Le type de l'étoile gouverne sa température, sa masse et son diamètre qui varient tous dans le même sens. Une étoile chaude est plus brillante aussi plus active qu'une étoile froide. Elle tend aussi à augmenter la température de la planète. Inversement une étoile de masse plus faible raréfie les transits par la révolution plus lente de ses planètes. Les types spectraux les plus favorables sont F,G,K,M.

Bordé et al. (2003) ont pris en compte tous ces facteurs pour préciser les bornes du domaine de détection. Il font quelques hypothèses supplémentaire sur les facteurs instrumentaux et les algorithmes de détection :

– Algorithme de détection. Afin de limiter la probabilité Gaussienne à une fausse alarme unique pour toute la mission, on contraint la profondeur

du transit à $\Delta F \geqslant 7\sigma$, σ étant l'écart-type de la courbe de lumière. L'accumulation de k transits améliore le signal proportionnellement à $\sqrt{k}$. La confirmation de périodicité exige au minimum trois transits en 150 jours.

– Les bruits pris en compte sont le bruit quantique, la variabilité stellaire, les bruits électroniques, le fond de lumière zodiacale, le bruit de dépointage (décrit dans un chapitre ultérieur), et la contamination par les PSFs d'étoiles de fond.

Compte tenu de ces éléments, et en se bornant aux étoiles naines de la séquence principale Corot sera adapté pour la détection de :

– Planètes terrestres ($r \sim 2R_\oplus$) autour d'étoiles de type K2 à M2, jusqu'à la magnitude 14-15,

– Planètes de classe Uranus ($r \simeq 3R_\oplus$) à partir du type G2 jusqu'à la $15^{ème}$ magnitude,

– Planètes géantes ($r \geqslant 3R_\oplus$) dès le type F7,

– au-delà de $r \geqslant 5R_\oplus$, la détectivité cesse de dépendre de l'étoile.

3.4.1 Nombre de détections attendues

Estimer les nombre de détection est un exercice hasardeux car on ne sait presque rien de la fréquence des systèmes planétaires. Corot va aider à lever un coin de voile en la matière. On peut néanmoins appliquer la détectivité précédente à une distribution réaliste de types et distances stellaires. Bordé et al. (2003) ont utilisé le modèle de Robin & Creze (1986), connu sous le nom de modèle de Besançon. Pour rester indépendants de toute probabilité de planète, ils calculent en supposant la présence d'une planète par étoile. Sa distance suit une loi uniforme à partir de 0.05 Unité Astronomique réduite (i.e ramenée à la luminosité du Soleil), qui est la plus petite distance de planète connue. Ces calculs ne doivent pas être interprétées comme une prédiction. Le résultat est présenté table 3.4.

Le modèle de Besançon peut s'avérer très différent de la réalité des champs Corot , mais l'inconnue majeure est la fréquence des exoplanètes.

TAB. 3.4 – *Détections intégrées pour une planète certaine/étoile/UA à partir de 0.05 UA. Valeur pour 60 000 étoiles naines. Il convient de retrancher 10% à cause de la contamination par les étoiles de fond. La valeur* $10R_\oplus$ *est indiquée car son taux de planètes est mieux connu et indique moins d'une planète par étoile. On se place dans l'hypothèse d'une variabilité stellaire d'amplitude Solaire.*

r ($R_\oplus$)	Nombre de détection intégré
1.0	5
1.25	12
1.5	26
2.0	70
3.0	189
5.0	300
(10.0)	311

Première partie

Fenêtrage optimisé

Chapitre 4

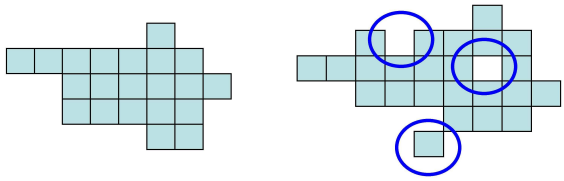
Énoncé des contraintes pour l'optimisation des masques photométriques

On est amené à définir une ouverture photométrique ou "fenêtre" pour chacune des 6 000 étoiles cible de chaque CCD, soit 12 000 au total, en limitant à 250 par CCD le nombre des différentes formes possibles. Ces formes sont appelées les "patrons". En effet, les 2×12 000 fenêtres ne peuvent utiliser que 2×250 patrons pour des raisons opérationnelles qui limitent le logiciel.

1. Le S/B d'une étoile munie d'un patron doit être le plus proche possible de celui qui est obtenu avec son ouverture taillée sur mesure. Cette ouverture spécifique est nommée "masque optimal".
2. Dans la mesure du possible, les patrons ne doivent pas être spécifiques d'un champ particulier, centre ou anticentre galactique, car il est préférable en terme de risque de conserver le même jeu de patrons pour les 5 champs de la mission.
3. Le jeu de patrons doit aussi se montrer robuste aux incertitudes sur le comportement de Corot en orbite, particulièrement pour ce qui concerne le jitter de pointage et la respiration jour/nuit.
4. Les ouvertures photométriques doivent respecter certaines limitations techniques dues à l'électronique de bord :

   – Une ouverture est un empilement de segments de ligne. Deux segments du même masque doivent être en contact par un pixel au minimum ce qui exclut les configurations non-connexes, les trous (voir Fig. 4.1),

   – l'intersection entre deux fenêtres doit être vide, même si les cibles sont proches,



a) b)

FIG. 4.1 – *a) Patron correct, b) Patron interdit. Les anomalies cerclées montrent de haut en bas un segment discontinu, un trou, l'absence de contact d'un pixel.*

– la surface du patron doit être inférieure à 150 pixels dans une fenêtre 25×20 pixels, et valoir en moyenne de 60 pixels pour limiter le temps de traitement à bord,

– il doit y avoir au plus 100 patrons qui coupent une ligne donnée,

– aucun patron ne doit chevaucher la colonne centrale d'un CCD, qui est inerte pour des raisons techniques, ni "mordre" sur les bords.

4.1 Terminologie

On nomme *masque* photométrique, l'ouverture virtuelle d'une cible à la surface du CCD. Il y a 12 000 masques de formes différentes. Chacun est optimisé pour une étoile de façon à donner un S/B aussi proche que possible de celui du bruit photonique pur.

On nomme *patrons* les formes en nombre restreint que peuvent prendre les masques. Il y à 250 patrons qui doivent préserver le mieux possible le S/B des 12 000 cibles.

On nomme *fenêtre* chacune des 12 000 ouvertures munie d'un des 250 patrons.

Le problème étant trop complexe pour être résolu globalement, on procède en trois étapes :

Etape 1 : Masques optimaux

A partir d'un grand nombre d'étoiles typiques de la variété des champs Corot , on crée autant de masques sur mesure, optimisés en termes de S/B . Ces étoiles "de travail" peuvent être en nombre quelconque, mais on choisi d'en utiliser 12 000 pour que les résultats soient représentatifs.

Etape 2 : Réduction

Les 12 000 masques optimaux sont réduits à 250 patrons. Nous avons adapté ou conçu quatre méthodes de réduction. Les trois premières visent à maximiser la ressemblance entre un masque optimal et son patron. La quatrième utilise directement le S/B comme métrique et précise la notion de S/B global, pour la relier aux dégradations individuelles subies par les étoiles. Un critère d'acceptabilité est défini et une matrice "d'acceptabilité" permet de choisir les patrons parmi les masques optimaux les plus fréquemment acceptables.

Etape 3 : Affectation

Les patrons sont répartis sur les étoiles d'un champ de cibles, qui n'est pas nécessairement celui utilisé pour la réduction. Il faut choisir le meilleur patron pour chaque étoile, les patrons ne devant pas se chevaucher. La mise en oeuvre d'une file d'attente avec priorité au S/B permet d'optimiser à la fois le S/B des cibles tout en éliminant le moins possible de candidats.

Ces étapes peuvent être optimisées indépendamment. Elle sont précédées d'une définition du rapport signal à bruit qui guide les optimisations.

4.2 Critère de qualité pour le fenêtrage

Le critère de qualité qui guide le fenêtrage a pour objectif la détection du plus petit transit possible dans une courbe de lumière. La qualité d'une courbe de lumière augmente avec l'information issue de l'étoile et diminue avec le bruit. On utilise habituellement un rapport signal à bruit (S/B) ou son inverse, la détectivité.

Le signal d'intérêt est le flux informatif f issu de l'étoile. Le bruit est lié à la fluctuation de la mesure autour de sa moyenne. Classiquement on l'assimile à σ, écart-type des points de mesure. Cette définition est suffisante bien qu'elle ne tienne pas compte des particularités du bruit, corrélation, fréquences etc...dont la connaissance facilitera le filtrage des courbes de lumières. Le S/B peut donc

s'écrire :

$$\text{S/B} = \frac{f}{\sigma}$$

Nous avons d'une part le S/B analytique calculable, et d'autre un paramètre physique, le rayon de la planète lié à $\Delta f/f$ par l'équation 2.5. Par construction S/B évolue dans le même sens que $f/\Delta f$, si bien qu'un masque conçu pour maximiser S/B minimisera du même coup le rayon de la planète détectable.

On peut lier $\Delta f/f$ et S/B par un indice de confiance en cherchant quel S/B donne moins d'une chance sur 100 pour que l'écart Δf d'un point soit le fruit du hasard. Pour une distribution Gaussienne de Δf, cela implique que $\Delta f \geqslant 2.65\sigma \simeq \sqrt{7}\sigma$ (voir Fig. 4.2). On réalise cette condition en moyennant les 7 poses que compte une heure. Une telle durée est acceptable car elle reste brève devant la durée d'un transit (3 heures au minimum). Le transit le plus bref serait réduit à 3 points quasi-certains. L'hypothèse Gaussienne pour le bruit est admissible car ses sources sont multiples, indépendantes et du même ordre. Avec Corot , certains bruits sont quasi-Gaussiens (bruit quantique), d'autres sont les résidus de correction de facteurs déterministes.

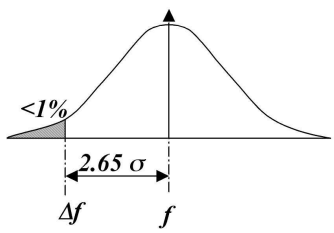


FIG. 4.2 – *Indice de confiance* 99% *à* -2.65σ*, pour une fonction de répartition Gaussienne.*

Le plus petit transit détectable à 99% de confiance par pixels est donc :

$$\boxed{\frac{\Delta f}{f}\text{détectable} = \frac{\sigma_{1\text{h}}}{f}}$$

Dans la suite, on parle indifféremment du S/B ou de son inverse quand le contexte n'est pas équivoque.

Chapitre 5

Calcul des masques optimaux

Il nous faut calculer le S/B un grand nombre de fois durant les phases de réduction et d'attribution. Dans cette section, je reprends les modèles de S/B analytiques que j'ai exposé dans la publication Llebaria et al. (2002) et les compare avec des simulations. Dans le modèle photonique contaminé, je fais une rapide estimation de l'effet du taux de contamination en fonction de la magnitude. Les petites lettres désignent les objets isolés (tel le flux f), et les grandes (F) leurs pendants contaminés (i.e contenant des photons qui n'appartiennent pas à la cible).

5.1 Le signal

Le signal choisi dans l'expression du S/B est le flux f de l'étoile. On pourrait penser qu'il est équivalent de choisir F, flux total comprenant l'étoile et le fond car la photométrie différentielle devrait retrancher d'elle-même la contamination. Il n'en est rien car l'utilisation de F tendrait à privilégier l'entrée de la contamination, par ailleurs bruitée.

Ce choix de f a comme conséquence que le rapport signal à bruit F/σ mesuré sur une courbe de lumière sera supérieur au S/B théorique car la courbe inclut des photons contaminants. Par exemple mesurer $\sigma = 1.0 \times 10^{-3}$ relatif sur une courbe, alors que le S/B prédisait $\sigma/f = 1.1 \times 10^{-3}$ ne signifie pas que la réalité est meilleure. Cela veut simplement dire que la courbe contient 10% de photons étrangers et qu'il faut une planète de plus grande dimension ($\Delta f/f = 1.1 \times 10^{-3}$) pour provoquer une baisse de flux égale au σ mesuré.

Il faut se rappeler que la contamination n'est pas facilement identifiable dans les courbes, sa mesure est possible seulement à partir d'images.

5.2 Les bruits

Ce sont les bruits qui déterminent la forme du masque. Par exemple ils en limitent l'extension en proscrivant les pixels trop éloignés qui apportent plus de bruit que de signal. Voici la liste des principaux bruits avec leurs caractéristiques.

5.2.1 Le bruit photonique

Nous avons vu que le bruit photonique d'un flux de moyenne f vaut

$$\sigma = \sqrt{n}$$

Donc le rapport signal à bruit purement photonique d'un flux vaut

$$\boxed{\frac{f}{\sigma} = \sqrt{f}}$$

Pour doubler le S/B , il faut donc quadrupler le flux. Corot à été conçu pour que la somme des autres bruits n'excède pas le bruit photonique.

5.2.2 La contamination

La contamination est la lumière provenant d'étoiles d'arrière plan, mesurées en même temps que la cible du fait du chevauchement des PSFs. A l'ordre 0 elle dépend de l'environnement de chaque cible. C'est un des principaux critères de sélection des cibles. Plus finement, elle dépend du masque de mesure et est donc prise en compte dans l'optimisation des masques. La contamination (notée c) agit par augmentation du bruit photonique dans la fenêtre. On mesure le taux de contamination d'une étoile de flux f par :

$$\tau = \frac{c}{f + c} \tag{5.1}$$

Une étoile contaminée à 50% compte autant de photons étrangers que de photons appartenant à la cible.

Il y a deux types de contamination : la contamination du fond continu et celle d'origine stellaire. Seule la deuxième dépend du choix des cibles.

La contamination stellaire utilisée pour le choix des cibles est estimée à partir d'images simulées en se dotant d'une PSF fixe, donc indépendante de la position de l'étoile. On la mesure à l'aide du masque optimal de la cible non contaminée. Nous ferons dans la suite l'hypothèse suivante : pour être retenue comme cible,

une étoile doit satisfaire $\tau \leq 10\%$. Les sources contaminantes sont par ordre d'importance :

Les étoiles proches. C'est la principale source de contamination. La PSF de la cible chevauche la PSF des étoiles de fond distantes de moins d'une dizaine de secondes d'arc. Les coordonnées des contaminantes sont connues, ces étoiles contraignent fortement la forme du masque. La procédure de sélection consistera à les exclure du masque le plus possible.

La lumière solaire rétro-diffusée par la terre. Malgré l'atténuation de 10^{-13} apportée par le baffle, une quantité non négligeable de photons diffusés parvient à s'immiscer dans le circuit optique. Cette lumière est réduite à 7 e^-/s/pixel, mais il peut subsister, suivant la position de la Terre et du Soleil, des pics plus importants. Ce contaminant n'est pas uniforme à l'échelle du CCD, mais est considéré comme tel à l'échelle d'un masque. Il influe peu sur la forme du masque.

La lumière zodiacale. C'est la lumière diffusée par les poussières interplanétaires situées dans le plan de l'écliptique. Suivant la ligne de visée, elle contribue pour environ 12 e^-/s/pixel. Elle varie annuellement de $\sim 10\%$.

Le fond continu des étoiles non résolues. Les étoiles plus faibles que la magnitude 21 ne sont pas cataloguées par les observations préparatoires au sol. En effet il faudrait effectuer une pose courte pour caractériser les étoiles fortes sans saturer le télescope et une pose longue sur le même champ pour déceler les étoiles faibles. Ce doublement du nombre de poses est trop exigeant en temps de télescope. De plus durant la pose longue on perdrait les étoiles faibles proches des cibles par saturation. Pour les besoins de la simulation, on assimile ces étoiles faibles et nombreuses à un fond continu supplémentaire de 3 e^-/s/pixel. Ce fond est estimé en extrapolant l'histogramme figure 3.5 au-delà de la magnitude de coupure du catalogue.

5.2.3 Le jitter

Les inévitables fluctuations résiduelles de pointage subsistant après l'asservissement ont pour effet de faire osciller l'image à l'intérieur du masque. Cette vibration de ~ 0.1 pixel d'amplitude se produit à un temps caractéristique de l'ordre de la seconde. Par conséquent des photons frontaliers sortent et entrent en permanence dans le masque engendrant une variation du signal lumineux comme illustré figure 5.1. Ce bruit de gigue nommé *jitter* est un facteur important pour

la forme du masque. Son rôle est étudié section 5.3. Le jitter est mesuré en permanence par le satellite, il sera en partie corrigé dans les données à l'aide d'un modèle de PFS.

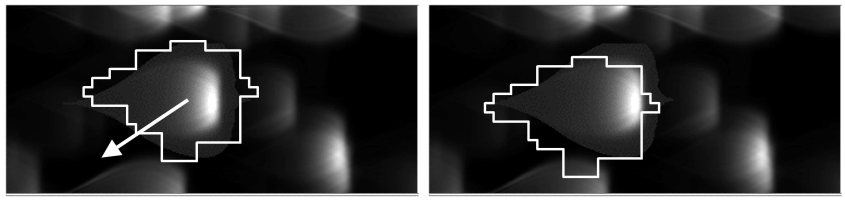


FIG. 5.1 – *Jitter. Les déplacements de PSF causent des entrées-sorties de flux. En quantités inégales, elles sont à l'origine du bruit de jitter.*

La signature du jitter est très sous-échantillonnée pour les temps de pose de 512 s. Les périodes impliquées sont donc réparties sur le spectre par le phénomène d'aliasing. On l'assimile à un bruit blanc.

5.2.4 La respiration

Le satellite parcourt une moitié de son orbite exposé au Soleil tandis qu'il reste dans l'ombre de la Terre durant l'autre. Sa température varie donc ; il s'ensuit des déformations susceptibles de modifier légèrement la focalisation du télescope. A nouveau, des photons vont franchir la frontière du masque induisant le bruit de *respiration*. On modélise cet effet par une dilatation-contraction centrée du masque, avec un temps caractéristique de 1h. On néglige l'interpénétration variable des PSFs et l'on considère que la dilatation reste centrée.

Le satellite est protégé par une couche de MLI (Multi Layer Isolation) à l'aspect de papier aluminium doré qui limite efficacement les transferts radiatifs. Les données actuelles laissent prévoir une amplitude de respiration de ~ 0.2 pixel.

L'effet de la respiration est une dérive périodique du flux moyen conjuguée à la variation de bruit photonique. On s'estime à même de corriger partiellement cette dérive dans les signaux reçus, notamment par filtrage ou à l'aide d'un repliement à la période orbitale pour déterminer un profil local. La respiration n'entre donc pas intégralement en ligne de compte dans la forme du masque. On la fait intervenir pondérée d'un coefficient résiduel reflétant l'erreur de correction et le supplément de bruit photonique.

5.2.5 Les bruits électroniques

Ces bruits apparaissent dans la partie analogique du trajet des électrons. Voici leurs principales causes :

Le bruit de lecture. La lecture du CCD se déroule en véhiculant les électrons de pixel en pixel le long d'une colonne. Puis la dernière ligne est décalée pixel à pixel jusqu'au convertisseur de sortie qui transforme la charge électrostatique en nombre entier. La polarité des puits de potentiels que sont les pixels est permutée pour pousser les électrons vers la sortie, pixel après pixel comme dans un jeu de Taquin. Certains électrons restent piégés dans les puits, et leur nombre variable engendre un bruit. Ce bruit qui reste inférieur à 0.5 e^-/s/pixel est négligé.

Le courant d'obscurité. L'agitation thermique crée spontanément des électrons qui sont piégés par les puits. Ils sont comptabilisés à tort comme photons. La variabilité de leur nombre est proportionnelle à $\sqrt{T}$, où T est la température en Kelvins. C'est pourquoi le CCD est maintenu à -40℃. Le bruit attendu est de 0.5 e^-/s/pixel en début de mission, augmentant avec le vieillissement jusqu'à 10 e^-/s/pixel.

Le bruit électronique proprement dit est inhérent au passage du courant dans tout circuit électronique. Il est blanc et proportionnel à $\sqrt{RT}$ où R est la résistance équivalente du circuit. Pour le limiter, les boîtiers électroniques sont soigneusement régulés en température. Avec le courant d'obscurité, ces bruits sont équivalents à une contamination uniforme supplémentaire de 1 e^-/s/pixel

Etant uniformes et faibles, les bruits électroniques n'interviennent que très peu dans la forme du masque.

5.2.6 Le traînage

Corot est dépourvu d'obturateur occultant le CCD. Pendant la lecture, il continue à recevoir les photons, créant des traînées lumineuses sur toute la colonne suivant le mécanisme expliqué figure 5.2. Chaque colonne reçoit une contamination uniforme égale à $t_{lec}.\sum_i f$ où $t_{lec} = 0.3$ s est le temps de lecture, et $\sum_i f$ est le flux sommé sur les i pixels de la colonne.

Ce phénomène est particulièrement problématique dans le cas des étoiles très brillantes, saturées ou non. Il étend à toute la colonne la portée de la contamination qu'elles induisent. Le traînage dépend de tout le champ. Par conséquent il n'est pas pris en compte dans le masque mais seulement lors de l'attribution des patrons.

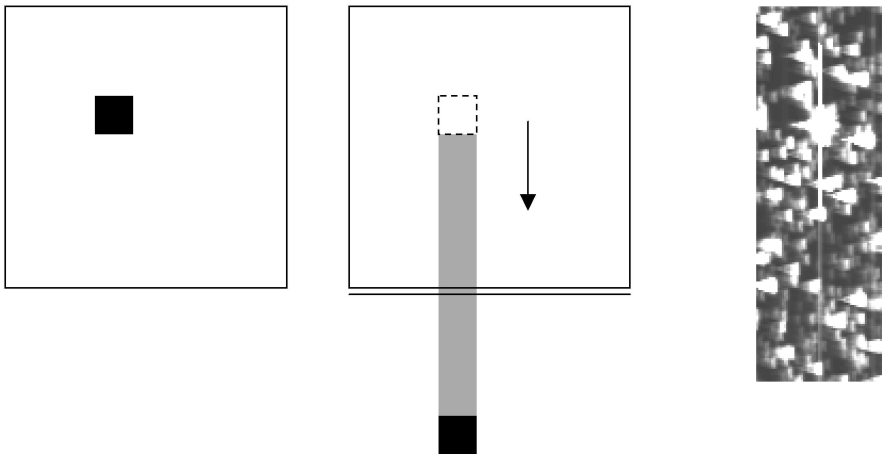


FIG. 5.2 – *Les photons de l'étoile (à gauche) continuent à arriver durant la lecture lors même que d'autres pixels passent sous l'étoile (au centre). La colonne entière s'en trouve éclairée. Certains photons seront comptabilisés dans d'autres pixels, ceux qui ne l'ont pas été le seront à la lecture suivante. A droite une image simulée, la saturation d'une étoile rend bien visible l'excès de traînage.*

5.2.7 La saturation

La saturation n'est pas un bruit à proprement parler mais elle pose des problèmes particuliers. Malgré la défocalisation, les étoiles plus brillantes que $m_V \leqslant 12$ vont saturer, plus particulièrement les pixels situés sous le maximum de la PSF. Ce phénomène n'est pas une simple redistribution des électrons, il rend les pixels impropres à la mesure. Les meilleures étoiles sont ainsi perdues. Heureusement elles sont rares. En revanche la saturation peut agir à longue portée par traînage (Fig. 5.2, droite) dégradant d'autres étoiles.

5.2.8 Bruits d'arrondi

Lors de l'analyse des données, la valeurs des quantités numériques telle le flux reste très éloignées des extrêmes manipulables par un ordinateur. Ceci limite les erreurs d'arrondi, mais on restera vigilants sur l'ordre de grandeur des résultats intermédiaires. Le problème est présent lors de la combinaison de deux nombres d'échelles éloignées, ou au contraire lors de la soustraction de deux nombres proches. Par exemple si $(x + \varepsilon)$ résulte d'un calcul arrondi à la précision δ négligeable devant x mais pas devant ε, la différence $(x + \varepsilon) - x = \varepsilon$ est instable car elle propulse δ au premier plan. On trouve une analyse intéressante du bruit numérique dans l'ouvrage de Press et al. (1997)

5.2.9 La variabilité stellaire

L'étoile cible est le siège d'une activité propre décrite dans la section 2.2.5. Provenant de la source d'information elle-même, ce bruit ne peut pas être amélioré par la forme du masque. Il peut être pris en compte en amont par un choix des cibles. L'activité stellaire sera traitée en aval lors de l'analyse des données reçues. Les masques ne prennent pas non plus en compte la variabilité des étoiles de fond.

5.2.10 Les éclipses d'étoiles du fond

On estime qu'une étoile de fond sur deux est susceptible d'être formée de deux étoiles non résolues gravitant l'une autour de l'autre. Pour une telle binaire 2000 fois plus faible que la cible, l'éclipse totale vue dans le plan de l'orbite commune produit une baisse de flux $\Delta F/F = 1/1000$, tout à fait analogue à un transit sur la cible. La magnitude d'une telle étoile vaudrait $m' = m + 6.7$, où m est la magnitude de la cible. D'après la relation constatée sur l'histogramme des magnitudes (voir Fig. 3.5), pour n étoiles de magnitude m on a :

$$n' \simeq n \times 2.2^{m'-m}$$

où n' est le nombre d'étoiles de magnitude m'. Numériquement elles sont 200 fois plus nombreuses. Heureusement toutes ne sont pas dans le masque et leur chance d'être vues dans le plan est limitée. Des simulations récentes semblent montrer que les binaires gênantes sont néanmoins en nombre significatif.

A l'heure actuelle, ce problème n'est pas pris ne compte dans la réalisation des masques. D'une part l'information de binarité d'une étoile n'est pas connue. D'autre diminuer le nombre d'étoiles d'arrière plan en réduisant la taille des ouvertures diminuer du même coup le flux de la cible et il faudrait définir des équivalences pour régler le compromis. De surcroît, l'effort fait pour exclure certaines étoiles d'arrière plan ne va pas dans le sens de la généricité des masques.

Certains de ces phénomènes seront distinguables par la différence de signature entre éclipse antérieure et postérieure, ou bien par un suivi au sol par la méthode des vitesses radiales.

Importance des différents Bruits

En pratique, il est spécifié que le cumul de toutes les dégradations ne doit pas excéder deux fois le bruit de photons. Corot est spécifié pour une détectivité $(\mathrm{S/B})^{-1} \leq 7 \times 10^{-4}$ pour $m_V \leq 15.5$, sur une heure.

5.3 Modèle du bruit photonique

Le calcul du S/B ne tient compte ici que du bruit photonique des étoiles résolues et des sources de contamination énumérées au §5.2.2. D'après l'expression de la variance d'un flux photonique, le S/B d'une étoile de flux f contaminée par un flux c vaut :

$$\text{S/B} = \frac{f}{\sqrt{f+c}} \tag{5.2}$$

C'est la limite physique que l'on cherche à atteindre. La table 5.1 indique quelques valeurs pour des magnitudes et contaminations types. A taux de contamination égal, l'influence des étoiles proches est constante, mais l'importance du fond uniforme augmente pour les étoiles faibles. Ceci reste vrai même si on réduit la dimension du masque.

TAB. 5.1 – *Limite de sensibilité. Pour chaque magnitude, on donne q et le rayon de la plus petite planète en fonction de trois contaminations : 1) l'étoile seule, 2)* $\tau = 10\%$ *d'étoiles proches et 3) ajout du fond uniforme* $c_0 = 13e^-/s/pixel$. *La taille du masque est adaptée à* f.

m_V	(e^-/s)	τ	q^{-1}(1h)	$r/R_\oplus$(1h)
12.0	24700	0	9429	1,1
		0.1	−5.1%	1,2
		$0.1 + c_0\times$ 100 pix	−7.3%	1,2
14.0	3912	0	3752	1,8
		0.1	−5.1%	1,8
		$0.1 + c_0\times$ 75 pix	−14%	1,9
15.5	978	0	1876	2,5
		0.1	−5.1%	2,6
		$0.1 + c_0\times$ 50 pix	−25%	2,9

La dégradation de S/B en fonction du taux de contamination τ s'obtient en substituant c (cf. Eq. 5.1) dans l'équation précédente. On trouve :

$$\text{S/B} = \sqrt{f}.\sqrt{1-\tau}$$

Les forts flux rendent tolérants à la contamination : à S/B fixé, une étoile brillante admettra un taux de contamination plus élevé qu'une étoile peu brillante.

5.4 Modèle jitter 1-D

L'approche à une dimension permet de comprendre l'interaction entre bruit photonique et jitter. La PSF φ répartit continûment les photons incidents à la surface du CCD. On voit figure 5.3 que le flux moyen f reçu dans un masque mono-dimensionnel $[X, Y]$ vaut :

$$f = \int_X^Y \varphi(x)\, dx \tag{5.3}$$

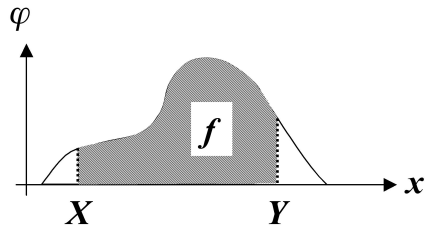


FIG. 5.3 – *Lien entre flux f et PSF φ*

Tout dépointage élémentaire λ causé par le jitter déplace les frontières, modifiant f :

$$f(\lambda) = \int_{X-\lambda}^{Y-\lambda} \varphi(x)\, dx = \int_X^Y \varphi(x-\lambda)\, dx \tag{5.4}$$

Etudions ce qu'il advient de q lorsqu'on fait varier λ de façon aléatoire. Les lettres majuscules telles F ou ϕ désignent les grandeurs contaminées.

5.4.1 Signal

Le signal est l'espérance mathématique $\langle f \rangle$ du flux "jitté" *non contaminé*. Si $\mathcal{P}(i)$ est la probabilité globale de recevoir exactement i photons, par définition :

$$\langle f \rangle = \sum_{i=0}^{\infty} i.\mathcal{P}(i) \tag{5.5}$$

On écrit $\mathcal{P}(i)$ en fonction de λ par la relation de Bayes sur les probabilités conditionnelles :

$$\mathcal{P}(i) = \int_{\lambda=-\infty}^{+\infty} \mathcal{P}_{|\lambda}(i)\mathcal{Q}(\lambda)\, d\lambda$$

où $\mathcal{Q}$ est la loi de probabilité du jitter supposée connue. On peut à présent substituer $\mathcal{P}_i$ dans l'équation 5.5. En permutant les signes $\sum$ et $\int$ il vient :

$$\langle f \rangle = \int_\lambda \underbrace{\left(\sum_i i\mathcal{P}_{|\lambda}(i) \right)}_{\left\langle f_{|\lambda} \right\rangle} \mathcal{Q}(\lambda)\, d\lambda \tag{5.6}$$

L'expression entre parenthèses n'est autre que le flux moyen pour un dépointage λ, c'est-à-dire le flux photonique $f(\lambda)$. Ce flux est donné par l'équation 5.4, ce qui amène :

$$\langle f \rangle = \int_\lambda \int_X^Y \varphi(x-\lambda)\mathcal{Q}(\lambda)\, dx\, d\lambda \tag{5.7}$$

En permutant les intégrales on reconnaît le produit de convolution :

$$\boxed{\langle f \rangle = \int_X^Y (\,\varphi * \mathcal{Q}\,)(x)\, dx} \tag{5.8}$$

Le signal moyen en présence de jitter est celui de la PSF de l'étoile, élargie par convolution avec le jitter.

5.4.2 Bruit au 1^{er} ordre

Nous allons calculer la variance du flux jitté F. Etant jitté, il ne suit pas forcément la loi Poisson, il faut repartir de la variance $\sigma^2 = \langle F^2 \rangle - \langle F \rangle^2$. Pour le premier terme, on revient à la définition :

$$\left\langle F^2 \right\rangle = \sum_{j=0}^{\infty} j^2.\mathcal{P}'(j)$$

Où $\mathcal{P}'$ est la loi de photon de l'étoile en présence de sa contamination et du jitter. En introduisant λ de la même manière que pour le signal on obtient :

$$\left\langle F^2 \right\rangle = \int_\lambda \underbrace{\left(\sum_j j^2 \mathcal{P}'_{|\lambda}(j) \right)}_{\left\langle F^2_{|\lambda} \right\rangle} \mathcal{Q}(\lambda)\, d\lambda$$

$\mathcal{P}'_{|\lambda}$ est la loi de probabilité des photons *non jittés*; c'est une loi de Poisson. Une propriété connue de cette loi concerne l'espérance du carré :

$$\left\langle F^2_{|\lambda} \right\rangle = \left\langle F_{|\lambda} \right\rangle^2 + \left\langle F_{|\lambda} \right\rangle$$

.

Le terme $\left\langle F_{|\lambda} \right\rangle$ est simplement le flux $F(\lambda)$. On peut donc écrire la variance :

$$\sigma^2 = \int_\lambda F(\lambda)^2 \mathcal{Q}(\lambda)\, d\lambda + \int_\lambda F(\lambda)\mathcal{Q}(\lambda)\, d\lambda - \left(\int_\lambda F(\lambda)\mathcal{Q}(\lambda)\, d\lambda \right)^2 \tag{5.9}$$

Il n'existe pas de relation simplificatrice entre $\int(\cdot^2)$ et $(\int \cdot)^2$ pour des fonctions quelconques. Intéressons nous donc à la tendance de σ pour de petits jitters en linéarisant F. Ceci est légitime car en pratique l'ordre du jitter sera inférieur à 0.2 pixel, petit devant les 10 pixels typiques d'un masque. Nous aurons soin plus loin de mesurer l'erreur commise à l'aide de simulations. Au 1^{er} ordre :

$$F(\lambda) \simeq F(0) + \lambda \frac{dF}{dx}(0)$$

On substitue cette expression dans l'équation 5.9. En simplifiant $\langle\lambda\rangle = 0$ et $\langle\lambda^2\rangle = \sigma_\lambda$, il ne reste que :

$$\sigma^2 = F(0) + \sigma_\lambda^2 \frac{dF}{dx}(0)$$

On peut expliciter $F(0)$ et $dF/dx(0)$ en revenant à la définition de la PSF :

$$\left\{ \begin{array}{lcl} F(0) & = & F_{\textit{non jitté}} \\ (dF/dx)(0) & = & \left(\int_{X-\lambda}^{Y-\lambda} \phi(x)\, dx \right)'(0) \quad = \quad \phi(Y) - \phi(X) \end{array} \right.$$

d'où la variance totale du flux jitté :

$$\sigma^2 = F_{\textit{non jitté}} + \sigma_\lambda^2 (\phi(Y) - \phi(X))^2$$

L'expression complète du S/B 1-D jitté au 1^{er} ordre est :

$$\boxed{q = \frac{f_{\varphi * \mathcal{Q}}}{\sqrt{F + \sigma_\lambda^2 (\phi(Y) - \phi(X))^2}}}$$

Le signal est le flux d'une PSF convoluée. On retrouve au dénominateur le terme $\sqrt{F}$ du bruit photonique contaminé, dégradé par un terme de jitter. Ce nouveau facteur est sensible à la différence de valeur de la PSF prise entre bords opposés du masque. Comme montre la Fig. 5.4, chaque déplacement élémentaire fait entrer et sortir des flux inégaux. Il faut donc privilégier les masques dessinés sur une ligne de niveau de la PSF, ou plutôt chercher à s'en approcher puisque

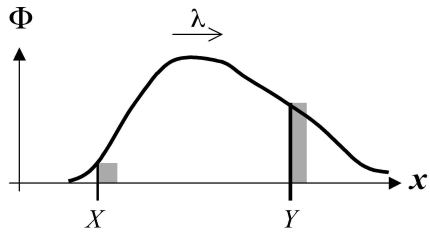


FIG. 5.4 – *La variation de flux lors d'un déplacement $d\lambda$ est la différence de hauteur entre les zones hachurées $\phi(X)d\lambda$ et $\phi(Y)d\lambda$.*

les pixels sont par nature discontinus.

La principale difficulté à laquelle on se heurte pour définir des masques photométriques se fait jour : un pixel individuel n'a pas de sens en termes de S/B car celui-ci dépend de toute la frontière délimitant le masque. Pour la surmonter on est donc contraint à de lourdes explorations parmi les combinaisons possibles. Les principaux ingrédients d'une optimisation efficace apparaissent :

1. Englober le plus de PSF de l'étoile au numérateur,
2. Exclure au mieux les photons contaminés du dénominateur,
3. Suivre une ligne de niveau, pour minimiser le jitter.

Cette question est détaillée par Llebaria (Llebaria et al. 2002).

5.4.3 Bruit au $2^{ème}$ ordre

Améliorons la précision en développant à présent $F(\lambda)$ au second ordre :

$$F(\lambda) \simeq F(0) + \lambda \frac{dF}{dx}(0) + \frac{\lambda^2}{2}\frac{d^2F}{dx^2}(0)$$

Le nouveau terme $(d^2F/dx^2)(0) = \phi'(X) - \phi'(Y)$ introduit une correction liée à la différence des pentes au bord du masque. Hélas, réintroduite dans l'équation 5.9 l'expression finale se simplifie peu, conservant des termes croisés avec des moments d'ordre 3 et 4 de la loi $\mathcal{Q}(\lambda)$. Cette expression n'est pas développée ici.

5.5 Modèle jitter 2-D

Les PSFs $\varphi(x, y)$, la loi de jitter $\mathcal{Q}(\lambda_x, \lambda_y)$ et le masque S sont en réalité des objets à deux dimensions. Nous allons généraliser le cas 1-D suivant les mêmes

étapes. Le signal (cf. Eq. 5.6) devient :

$$\langle f \rangle = \iint_{\lambda_x,\lambda_y} \left(\sum_i i\mathcal{P}_{|\lambda_x,\lambda_y}(i) \right) \mathcal{Q}(\lambda_x,\lambda_y)\, d\lambda_x d\lambda_y$$

On reconnaît entre parenthèses le flux photonique à travers une PSF déplacée de (λ_x, λ_y) :

$$\langle f \rangle = \iint_{\lambda_x,\lambda_y} \oint_S \varphi(x-\lambda_x, y-\lambda_y)\, dxdy \mathcal{Q}(\lambda_x,\lambda_y)\, d\lambda_x d\lambda_y$$

En permutant l'ordre de ces intégrales on obtient le produit de convolution 2-D étendu à toute la surface du masque :

$$\boxed{\langle f \rangle = \oint_S (\,\varphi * \mathcal{Q}\,)(x,y)\, dxdy} \tag{5.10}$$

5.5.1 Bruit 2-D au 1^{er} ordre

Pour calculer la variance du flux, on commence cette fois-ci par le linéariser au premier ordre en fonction d'une perturbation de pointage $\vec{\lambda}$:

$$F(\vec{\lambda}) \simeq F(\vec{0}) + \vec{\lambda}^t.\overrightarrow{\text{grad}}F(\vec{0})$$

Par additivité des variances :

$$\text{var}F(\vec{\lambda}) = \text{var}F(\vec{0}) + \text{var}(\vec{\lambda}^t.\overrightarrow{\text{grad}}F(0))$$

D'une part $F(\vec{0})$ est le flux *non jitté*, sa variance de Poisson vaut donc F. D'autre part en tant que fluctuation $\vec{\lambda}$ est centrée, donc $\text{var}(\vec{\lambda}^t.\overrightarrow{\text{grad}}F) = \langle(\lambda_x.g_x + \lambda_y.g_y)^2\rangle$ où g_x et g_y sont les coordonnées du gradient. On obtient :

$$\text{var}F = F_{non\ jitté} + \sigma_x^2 g_x^2 + \sigma_y^2 g_y^2 + 2g_x g_y \text{cov}(\lambda_x,\lambda_y)$$

Explicitons g_x et g_y. Par définition de la PSF :

$$\begin{aligned}
\left[\overrightarrow{\text{grad}}F\right]_{\substack{\lambda_x=0\\ \lambda_y=0}} &= \overrightarrow{\text{grad}}\left[\oint_S \phi(x-\lambda_x, y-\lambda_y)\, dxdy\right]_{\substack{\lambda_x=0\\ \lambda_y=0}} \\
&= \oint_S \left[\overrightarrow{\text{grad}}\phi\right]_{\substack{\lambda_x=0\\ \lambda_y=0}} dxdy \\
&= \oint_C \vec{n}\phi(x,y)\, dxdy
\end{aligned}$$

Où $\vec{n}$ est la normale unitaire qui parcourt le contour C du masque. Le gradient est bien un vecteur à deux coordonnées. Dans notre cas, la décomposition sur x, y est très simple car C suit des pixels carrés. Il en découle :

$$\left\{ \begin{array}{rcccl} g_x & = & \sum \phi(x') - \phi(x) & = & \Delta_x \\ g_y & = & \sum \phi(y') - \phi(y') & = & \Delta_y \end{array} \right.$$

Les x, x', y, y' sont les bords gauche, droit, inférieur et supérieur du masque, Δ_x et Δ_y les différences de hauteur cumulée entre frontières opposées. Avec le numérateur, l'expression complète du rapport signal à bruit 2-D est :

$$\boxed{q = \frac{\oint_S \varphi * \mathcal{Q}}{F_{non\ jitté} + \sigma_x^2 \Delta_x^2 + \sigma_y^2 \Delta_y^2 + 2\mathrm{cov}(\lambda_x, \lambda_y)\Delta_x \Delta_y}}$$

Comparaison au modèle 1-D

On retrouve les mêmes éléments : La PSF convoluée au numérateur et la PSF contaminée au dénominateur. Elle n'intervient que par sa valeur aux frontières du masque. Les frontières horizontales et verticales sont sommées indépendamment comme si elles étaient contiguës et alignées. La dépendance entre λ_x et λ_y apparaît dans le terme additif de couplage $\mathrm{cov}(\lambda_x, \lambda_y)$. Pour comprendre le phénomène, supposons de petits déplacements λ_x et λ_y qui induisent respectivement des suppléments de flux df_x etdf_y. Le flux total devient plus erratique quand le couplage x, y augmente car alors df_x et df_y se somment sans plus pouvoir se compenser.

5.5.2 Bruit 2-D au $2^{ème}$ ordre

Nous avons étendu le calcul pour de plus grands jitters en développant F au $2^{ème}$ ordre. Pour présenter le résultat nous avons besoin des notations supplémentaires :

$$\left\{ \begin{array}{rcl} \Delta'_x & = & \int_y \vec{i}.\overrightarrow{\mathrm{grad}}\phi \, dy \\ \Delta'_y & = & \int_x \vec{j}.\overrightarrow{\mathrm{grad}}\phi \, dx \end{array} \right.$$

qui sont les différences cumulées de pente entre bords opposés du masque. Les moments d'ordre supérieur du jitter interviennent, mais l'expressions reste trop lourde à manipuler. Sous l'hypothèse simplificatrice d'indépendance des λ_x et λ_y, il en reste :

$$\begin{array}{rcl} \sigma^2 & = & F_{non\ jitté} + \sigma_x^2(\Delta_x^2 + \Delta'_x/2) + \sigma_y^2(\Delta_y^2 + \Delta'_y/2) + \\ & + & \sigma_x^4 {\Delta'_x}^2 + \sigma_y^4 {\Delta'_y}^2 + \sigma_x^2 \sigma_y^2 \Delta'_x \Delta'_y \end{array}$$

Les simulations du paragraphe suivant montrent que cette expression n'apporte pas de précision supplémentaire. En effet l'écart entre résultats des expressions au 1^{er} et au $2^{ème}$ ordre est négligeable devant l'écart entre formules analytiques et simulation.

5.6 Vérification expérimentale

J'ai comparé les S/B analytiques précédents avec un S/B obtenu par simulation. Nous verrons que c'est l'expression en 1-D du §5.4.2 qui assure le meilleur compromis précision-simplicité sous l'hypothèse de λ_x, λ_y indépendants. Le mode opératoire est le suivant :

- Un programme déplace aléatoirement une PSF sous un masque par pas de 1 seconde,
- Après chaque déplacement, on mesure le flux dans le masque,
- Les mesures sommées sur 512s constituent la courbe de lumière (voir Fig.5.5)
- On mesure dans la courbe le S/B simulé, en tenant compte de la contamination.
- On compare avec celui que l'on a calculé analytiquement avec les mêmes paramètres de jitter, PSF, ...

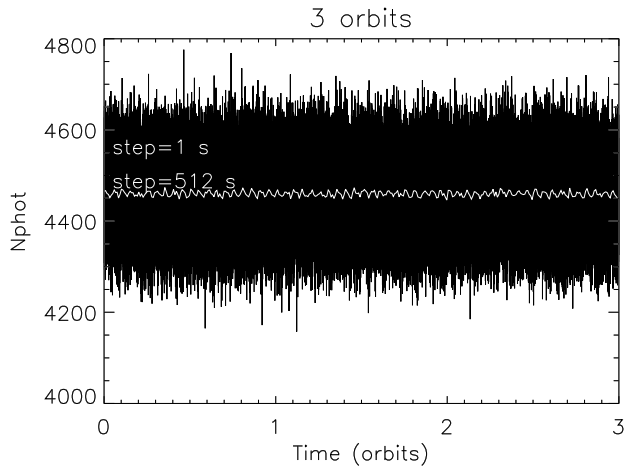


FIG. 5.5 – *Simulation de flux sur une seconde (en noir) et intégré sur 512 secondes (en blanc), pour* $m_V = 14$.

Le test a été effectué pour trois cibles $m_V = 11, 13$ et 14, contaminées à $\tau = 1.6, 3.9$ et 5.8%. Pour simuler le bruit de photons, les flux suivent une loi

Gaussienne $\mathrm{G}(n, \sqrt{n})$ avec le nombre de photons n adéquat. Les flux cible et contamination sont simulés séparément. En sortie, le S/B est estimé graphiquement par le rapport :

$$\widetilde{\mathrm{S/B}} = \frac{\overline{f}}{\widetilde{\sigma_F}}$$

La variation de pointage simulée est la somme de deux composantes. L'une aléatoire et Gaussienne autour du pointage moyen l'autre, déterministe, est le déplacement de ce pointage moyen au cours de l'orbite en fonction des harmoniques 2 et 4. Pour x par exemple :

$$\lambda_x(t) = \mathrm{G}(0, \sigma) + k_1 \sin(4\pi t/T_{\mathrm{orb}}) + k_2 \sin(8\pi t/T_{\mathrm{orb}}).$$

où σ est l'écart-type, k_1 et k_2 sont les coefficients des harmoniques 1 et 2. Le dépointage résultant en x, y est indiqué figure 5.6.

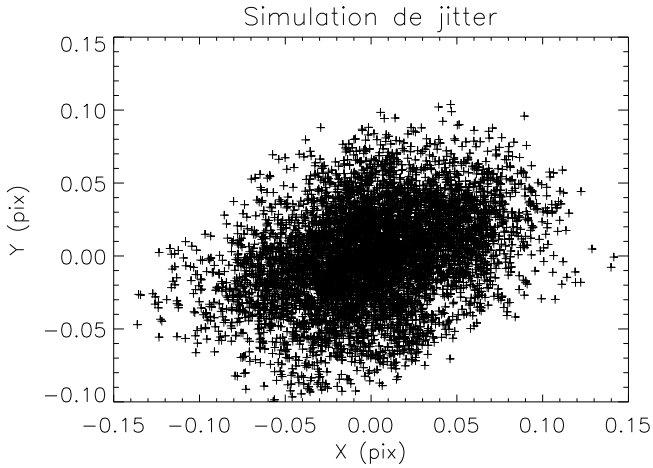


FIG. 5.6 – *Variations de pointage simulées sur trois orbites. Les coefficients du jitter sont* $\sigma = 0.18$*", *$k_1 = 0.15$*" et *$k_2 = 0.06$*" suivant les x et $\sigma = 0.16$*", *$k_1 = 0.12$*", *$k_2 = 0.05$*" suivant y.*

On compare les trois expressions de la section précédente :
- Le modèle 1-D, en additionnant le bruit indépendamment $\mathrm{var}\lambda = \mathrm{var}\lambda_x + \mathrm{var}\lambda_y$
- Le modèle 2-D au 1^{er} ordre, équivalant au précédent, corrigé par le couplage λ_x, λ_y mesuré,
- Le modèle 2-D au $2^{ème}$ ordre, similaire au second mais prenant en compte tous les moments d'ordre supérieurs.

La loi $\mathcal{Q}$ du jitter est estimée avec les points disponibles. Le produit de convolution est donné par le signal analytique :

$$s = \iint_{\lambda_x,\lambda_y} \oint_S \varphi(x-\lambda_x, y-\lambda_y)\, dxdy\mathcal{Q}(\lambda_x,\lambda_y)\, d\lambda_x d\lambda_y$$

se résume à

$$s = \frac{1}{n} \sum_{(\lambda_x,\lambda_y)} \varphi(x-\lambda_x, y-\lambda_y)m$$

où m est le masque binaire. La comparaison des trois modèles est présentée table 5.2.

TAB. 5.2 – *Comparaison entre S/B mesuré et formules analytiques. La simulation est conduite sur une durée de trois orbites. Les paramètres du jitter sont ceux de la figure précédente. Les taux de contamination sont indiqués entre parenthèses.*

m_V	$11.3(\tau = 1.6\%)$	$13.2(\tau = 3.9\%)$	$13.9(\tau = 5.8\%)$
simulé	1 463	784	1 583
1-D	2 404	482	2 924
2-D 1^{er} ordre	2 646	840	2 779
2-D $2^{\text{ème}}$ ordre	2 625	830	2 766

La non prise en compte de la covariance entre x et y dans l'expression 1-D entraîne une différence notable de résultat avec les expressions 2-D. Ces dernières au 1^{er} et au $2^{ème}$ ordre donnent des résultats proches ($\Delta \sim 1\%$), on retient donc l'expression 2-D au 1^{er} ordre en vertu de sa simplicité, bien qu'il subsiste une différence par rapport aux simulations.

5.7 Simulation d'Images

Avant le vol, il nous faut faire des simulations aussi réalistes que possible pour pallier l'absence d'images réelles. Ces images ont l'avantage de contenir plus d'information que des images réelles Corot : Elle gardent trace de l'identité et des proportions des sources éclairant un pixel donné. La simulation des images est décrite en détail dans l'article de Llebaria (Llebaria et al. 2002) adossé ci-après et rappelée ici.

Le point de départ de la simulation est :

- Une liste d'étoile extraite de la base EXODAT pour le pointage choisi. Cette liste contient notamment les magnitudes des étoiles du champ dans les bandes B,V,r,i,

- les PSFs de référence calculées pour 16 types-spectraux de références indexés par T, température de la photosphère et pour 9 positions par CCD. Ces PSFs sont échantillonnées au $1/5^{e}$ de la taille du pixel.

- la correspondance entre coordonnées angulaires de l'étoile et position (x, y) sur le CCD du point de référence de la PSF.

Le CCD initial est vierge. Puis la PSF de chacune des 50 000 à 250 000 étoiles du champ, cibles et non cibles, est simulée et accumulée à l'état courant du CCD.

5.7.1 Simulation des PSFs stellaires

Chaque cible est placée au centre d'une petite imagette de 40×26 pixels. Le point de départ est l'indice de couleur V-R$= m_V - m_R$ de la cible (magnitudes dans les bandes vertes et rouges). Il est indépendant de la magnitude de l'étoile et relativement peu sensible à l'absorption par le milieu interstellaire. Il est caractéristique du type spectral de l'étoile et l'on en déduit la température T de sa photosphère.

On extrait parmi les PSFs de référence celles encadrant T (choisies à la bonne position sur le CCD, voir figure 5.7) et l'on interpole leurs pixels pour obtenir la PSF polychromatique à normaliser par le flux total. La figure 5.8 donne quelques exemples de PSF pour différentes températures.

Il reste à positionner correctement sur le CCD cette PSF polychromatique fournie sur-échantillonnée au $1/5^{e}$ de pixel par ZEMAX, grâce à son point de référence (cf. §3.2.5), et à la re-échantillonner au pas du pixel comme la verra Corot .

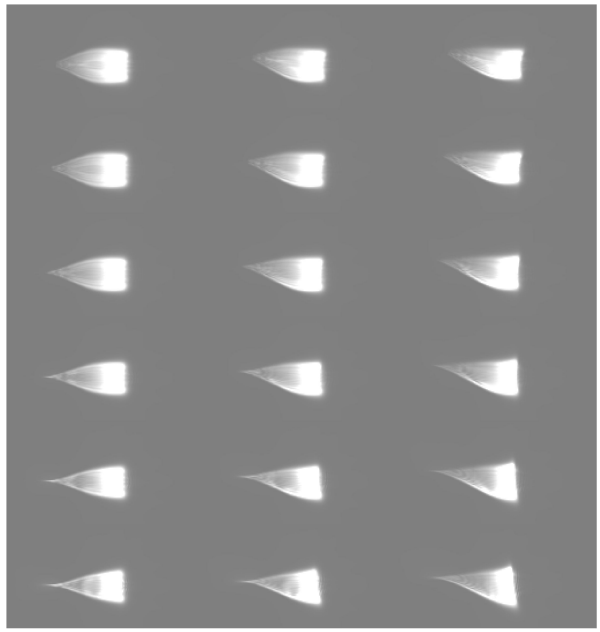


FIG. 5.7 – *La PSF d'une étoile de type G2, calculée en neuf positions différentes pour chaque CCD. Pour la simulation de PSFs stellaires on se contente de choisir un cadran plutôt que de faire appel à une interpolation 2D.*

La PSF de la cible ainsi calculée, on réitère avec les étoiles de fond présentes dans la trame de travail. Toutes les étoiles sont ainsi simulées ; les plus faibles seront noyées dans le bruit de fond mais néanmoins présentes et susceptibles d'engendrer des éclipses d'étoile double gênantes. Le catalogue disponible est cependant incomplet, la magnitude de coupure se situant vers $m_V \geqslant 19.5$ (voir Fig. 3.5). Les étoiles non résolues contribuent malgré tout à la PSF. On en tient compte sous forme d'un fond uniforme, dont le flux est calculé en intégrant la droite d'extrapolation figure 5.9.

Pour compléter, il est nécessaire d'ajouter au fond continu la lumière zodiacale et rétro-diffusée, ainsi que le courant d'obscurité et les autres bruits électroniques. On termine par les artefacts instrumentaux globaux : ailes de saturation et traînage. La figure 5.10 montre un exemple d'image Corot simulée.

Nous appelons *champ local* l'imagette d'une cible et celle de sa contamination associée. Un tel champ est conservé car il contient plus d'information qu'une image réelle (i.e) la séparation entre photons des cibles et photons contaminants. Malgré la multiplicité des opérations qu'il est nécessaire d'effectuer pour obtenir

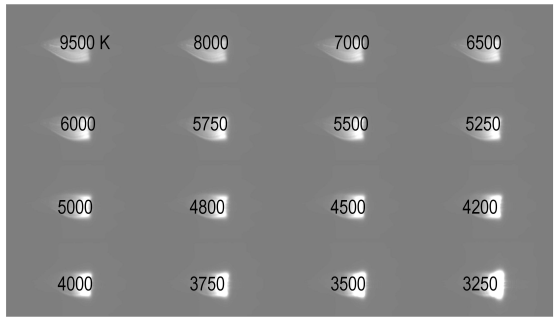


FIG. 5.8 – *PSF polychromatique pour différents types d'étoiles, résumés à la température en Kelvin de leur photosphère. A cause de la dispersion par le prisme le maximum de flux se décale vers le rouge (en bas à droite) quand la température de l'étoile décroît.*

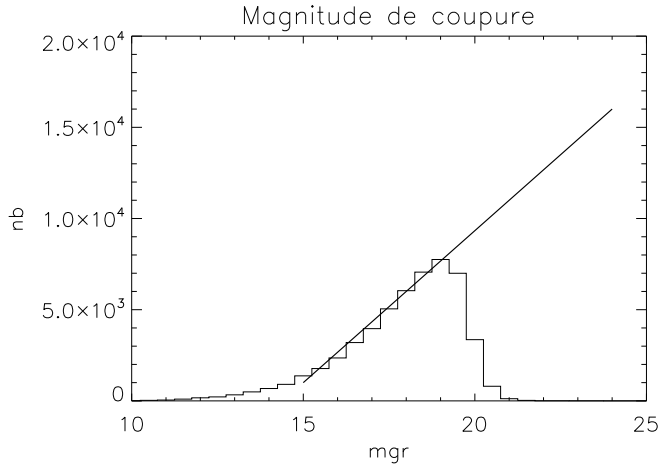


FIG. 5.9 – *Magnitude de coupure et extrapolation.*

une imagette, la simulation du champ complet ne dure guère plus de 1/2 heure.

5.7.2 Masques optimaux

La méthode de détermination de ces masques est décrite dans l'article de Llebaria (Llebaria et al. 2002). Les masques ayant une surface de 50 à 120 pixels, le nombre de possibilités est bien trop grand pour faire une recherche exhaustive. La procédure comprend deux étapes.

1. L'étape d'ébauche calcule le meilleur masque pour la cible et ses contaminants en l'absence du bruit de jitter et de respiration. Dans ce cas l'expression du S/B n'est pas liée à la forme de la frontière (cf. §5.3) ; On procède par classement : les pixels sont englobés dans le masque jusqu'à ce que le

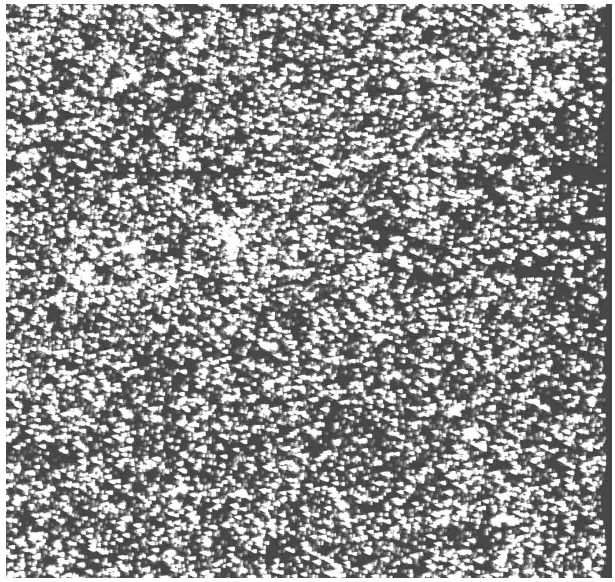


FIG. 5.10 – *Extrait de champ simulé. Cette image correspond à 1/10^e du champ vu par Corot . On remarque bien la densité du champ et le chevauchement des PSFs*

S/B cesse d'augmenter. L'algorithme travaille ligne par ligne pour garantir que tout pixel soit rattaché au masque par un côté au minimum.

2. La phase de finition est un ajustement en présence des effets de jitter et de respiration. L'expression du S/B dépend cette fois de la géométrie de la ligne frontière. Elle ne permet pas d'isoler le rôle d'un pixel donné : il faut tester toutes les formes de frontière possibles. En supposant la frontière idéale proche de la frontière ébauchée, on déforme cette dernière en testant une centaine de milliers de combinaisons. Le masque de meilleur S/B devient le masque optimal.

Cette opération est répétée pour chacune des 12 000 cibles.

Le S/B utilisé dans l'optimisation des masques doit tenir compte du vrai jitter $\sigma_\lambda \simeq 0.1$ pixel et non pas d'une valeur résiduelle après correction. En effet, une valeur trop petite produirait l'effet inverse de celui recherché : les courbes de lumière ne seraient pas optimisées en jitter, rendant moins efficace la correction. C'est seulement à l'issue du processus de fenêtrage que l'on peut mesurer les performances attendues après correction du jitter, en utilisant cette fois un coefficient résiduel ($\sigma_\lambda \simeq 0.02$) pour simuler une correction imparfaite.

5.8 Première publication Llebaria et al. (2002), SPIE.

Designing photometric patterns for exoplanet transit search on board COROT

Llebaria A. [a], Vuillemin A.[a], Guterman P.[b] and Barge P. [a]

[a]Laboratoire Astrophysique de Marseille(CNRS), 13776 BP8 Marseille Cedex12, France
[b]Gemplus, Aubagne, France

ABSTRACT

COROT is a mission of the CNES space agency, to be launched in 2005 in a near Polar orbit. It is devoted to star seismology and to exoplanetary transit search. Five star fields chosen close to the galactic plane will be observed during the mission with a high photometric accuracy (relative). Each observation run will last 150 days monitoring continuously more than 6000 stars. This paper presents a new method designed to perform optimal aperture photometry on board in high density fields. We describe the way the photometric windows or patterns are defined and centered on the CCD around each target star, with the expected performances. Each pattern depends on the specific 2D profile of the point spread function (PSF) but also on the pointing jitter and on the tiny deformations of the telescopes. These patterns will be stored on board in order to define for each target star the optimal pattern which will produce the integrated flux to be measured. This method allows a significant increase of the sampling rate to aproximatively one measure per star each 8 mn).

Keywords: image processing, exoplanets, pattern analysis, stellar photometry

1. INTRODUCTION

COROT (COnvection) ROtation and planetary Transits) is a mission developped in the framework of the CNES* small satellite programme with a wide european cooperation.[1,2] It has to be launched in 2005. The aim of COROT mission is double: 1) to monitor the sysmology of stars and 2) detect and measure exoplanetary transits.

1.1. Mission profile for exoplanets research

COROT will be located in a PROTEUS platform designed for low-Earth orbits. For long uninterrupted observations an inertial polar orbit is needed. The orbital plane can be chosen freely and will remain the same during all the mission.

Necessity to keep a very low level of straylight avoiding the Earth limb in the field of view, limits the visible sky of Corot to two circles of approximatively 10 to 15 degrees of radius, centered on the celestial equator. The position of this circles has been chosen to optimize the two scientific programmes: their centers are at right ascension of 102.5 ± 2 and 282.5 deg (where the equatorial plane crosses the galactic plane). The field of view is $3^o.46$ wide.

Yearly the observation time is divided in two periods; each period includes a long run (150 days) and two short runs (10 to 20 days). During each run COROT will measure simultaneously 6000 stars with a record rate of $\sim$ 6 measures/star/hour. The expected transits of planetary bodies will look like tiny notches in the photometric records of stars. The notch will be on the 3.10^{-4} range of the full signal and will last for some hours (between 3h and 10h is the work hypothesis). The relative precision of such measures, not their absolute accuracy, is the crucial point of this mission. Therefore the stability of experimental conditions and a very low straylight level are important items.

Further author information: (Send correspondence to Antoine Llebaria)
A.LL.: E-mail: Antoine.Llebaria@obmp.fr, Telephone: +33 (0)4 91 05 59 00
*the Centre National d'Etudes Spatiales (CNES) is the french spatial agency

1.2. Optical instrumentation

The optical scheme has been built to reduce the internal straylight and therefore the minimize the influence of the periodic changes on the illumination from the Sun and from the Earth. The optical instrument is an off axis telescope with an entrance pupil of 600 cm^2 and it includes a 2 mirrors afocal section and a 6 lenses dioptic objective, with 30 cm of diameter aperture and 1200 mm of focus, working in 370-950 nm range. This design was chosen because it is compact and it shows an hight rejection rate for straylight (better than 10^{13}.

The focal plane includes 4 CCDs of 2048×2048 pixels each (Marconi MPP), of 13.5 microns of pixel interval, working in frame transfert. Two are dedicated to seismology and the other two to exoplanetary detection. These ones work in focus but with a small dispersion (R$\sim$4) parallel to the rows of the CCD. Such dispersion is caused with a low dispersion prism which intercepts the optical path on the exoplanetary side in order to obtain some color information about the measured stars and therefore to get additionnal criteria to discriminate the transits from the otherwise normal stellar activity. So the best possible couloured PSF in the exoplanet field has a strong peak in the red and is more dispersed in the blue.

1.3. On board processing

Telemetry constraints (TM rate < 550 Mbits/day) force to on board heavy processing, to extract the photometric imformation from the images 2048×2048 pixels wide. In order to avoid the saturation of CCDs the exposure time is 32 sec. Each image is read in the auxiliary memory where the "extractor" (a specific programmable device) defines 6000 windows i.e. predeteminated subsets of adjacent pixels. Each window will cover a star image, therefore the number of selected stars define the total number of windows. The total flux and the red flux is measured in each window. For each of them the DPU (digital processing unit) adds 16 successive measures and each 8 minutes packs the final results in the telemetry (TM) format. A very small subset of windows are included in the TM flow as mini-images.

The DPU software allocates a particular window design (called pattern thereafter) to each selected star. All the windows are defined from the limited collection of 256 patterns preloaded from the ground. Loading new patterns from ground on the DPU is a very slow operation, is by far more fast to reload new positions only and to redirect the preloaded patterns to the new subset of stars. Therefore, when the star field changes, the extractor uses the same collection of patterns for the new set of positions.

The description of the pattern build-up process and its expected performances are the main objective of this paper. The experiment objectives and constraints are resumed on section 2. Simulated images are needed to define the set of patterns, their characteristics and build-up are detailed in section 3. Deduction of specific patterns for each star is the subject of section 4 and their redution on a small collection of generic masks is the subject of section 5. Expected performances are discussed in section 6.

2. PHOTOMETRY

Star photometry in the COROT fields is difficult because they are dense and the flux of each star is contaminated with the flux of neighbour stars. Crawded fields are usually reduced with *PSF fitting* technics, however in our case such processing methods are useless due to the operational constraints like DPU power, PSF variability, etc.. *Flux integration* was chosen here because it is fast enough and can be easely taylored to the variability of PSF aspect along the field and with the spectral type of the star. The prism in the optical pathway of the exo-planets channel disperses the star light over the CCD surface. For any star the final PSF image is the addition of a continuous set of monochrome PSFs in the useful wavelength range, changing in position and of form. They are weighted by the star spectral profile and the spectral responsivity of COROT which includes the CCD responsivity.

One of the most classical *flux integration* methods adds the flux arriving to two concentric windows of increasing diameter centered on the star image. Background is estimated by difference between both windows. Is possible to use a window alone if background can be estimated independently. The COROT primary purpose is to obtain stable measures from frame to frame, that is, only the relative precision between successive measures of each star matters, therefore only a window per star is needed. The background time stability can be monitored in *ad hoc* windows.

Optimal windows will include as much flux as possible comming from the star and as few flux as possible comming from its sourroundings, S/N has to be maximized. Aside any prerequisite, any pixel able to increase the total S/N has to be included in the window. S/N is weakened by platform dependent noises like jitter and breathing and by extrinsic noises due to the background level (generating Poisson noise), variable star neighbours, cosmic rays, stray light, erratic dust particles, etc.

Jitter is the residual movement of the line of sight after platform stabilization. The main star image is formed in one of the sysmology CCDs defining the line of sight, therefore there is a local movement relative to the CCD frames in each star image. The local movement, for each star, depends on the residual movement of the main star and on the small rigid body rotation around this target. Jitter is directly linked to attitude stability. Jitter is a wide-band phenomena relatively to the measure rate of 1 measure each 32 sec.

Breathing results from internal changes on the telescope structure due to the thermal chages and aging. It is synchonous with the orbit, but their effect is deeper than the effect of jitter because breathing modifies the PSF aspect. Due to the abrupt edges of the pattern, integration in windows will be very sensitive to jitter and breathing with variable backgrounds. The optimization procedure tries to minimize this crucial drawback.

Because the extractor device has a limited capacity, it is necessary to resume the full set of optimized windows in a limited collection of 256 patterns. These patterns are called generic masks. Each star image to be measured will be associated to a window which pattern has the "best" match with its optimized window. Consequently, photometric windows will be build-up in a multistep process.

- Image simulation.
- Defining the optimal window for each target.
- Resuming all patterns in a reduced collection of 256 generic masks.
- Bestowing each target upon one of generic masks.

Summarizig, the image of each mesurable star, the image of its background and the expected errors due to jitter and breathing are the basic elements to determine a window and its expected performance

3. STAR FIELD SIMULATION

Simulation of the future images is the only way we have to define the patterns. They are of two classes: the full CCD images and the local images. Only local images are useful here, full CCD images will be used for the test of the overall process including star detection and adjustement of processing parameters. Both simulations require many basic elements:

1. The catalog of stars on the field including positions, color index and spectral classes.
2. Background level including zodiacal light and straylight.
3. The collection of 2D PSFs depending on position over the CCD, spectral type and class.
4. The geometric parameters needed to determine the star images on the CCD plane.
5. The radiometric response of COROT

3.1. Star catalogs

A photometric study of the selected fields until the 19 mag(R) for the BVRI colours is in progress. In the meantime we use the DENIS and USNO catalogs as simulation basis. The histogram shows clearly the limits of both catalogs; their completude falls beyond the 19 mag(R). Therefore local number of stars has been statisticaly extrapolated to simulate the background beyond this magnitude.

The catalog provides: star positions, magnitudes, a set of color indexes and rates of measure reliability (only for a subset of these magnitudes). The mag(R) is used as reference because the maximum of COROT sensivity yelds in this domain of wavelenght, it is the most common measured magnitude and it is known for an overwhelming majority of cataloged stars.

3.2. PSFs

Local and global simulations use a large collection of modelized PSFs. Such PSFs depend mainly on the spectral type and class of the star and on the field position. The PSF collection is formed in successive steps:

1. Collecting a set of spectra in the $350 \ldots 1050$ nm wavelength range for a set of surface temperatures (correlative to spectral type) of the main sequence stars.
2. Modelling a very large set ($> 200 \times 18$) of monochrome PSFs for a group of 18 positions on the field.
3. Building up the collection of polychrome PSFs ($\sim 10 \times 18$) from the monochrome PSF applying the star spectral models, the CCDs response courbe and the COROT passband efficiency.

The PSF models have been built using optical design software packages (ASAP and ZEEMAX). Differences between theirs results are irrelevant. They show equivalent performances of run time.

The PSF choice for a specific star on the catalog uses its apparent color index aCI [†]. (if it is known). With any assumption about the interstellar reddenning, aCI is associated to an apparent surface temperature aT [‡]. i.e. the key value to chose the spectral profile.

3.3. Geometric parameters

Geometric parameters link a sight direction with a PSF position on the CCD. This parameters include external parameters like the COROT attitude, and internal parameters like mean focus distance and distortion. The reference is defined for 500 nm. of wavelenght. In summary:

- External parameters tie the star coordinates to COROT coordinates
- Internal parameters tie any sight directions in COROT coordinates to its position in CCDs plane (in mm) for the 500 nm wavelength.
- CCD parameters tie the each CCD positions to image positions.

This chain of references defines for each star on the field the position of its correlative image in the CCD image.

3.4. Simulated images

For local simulations the oversampled PSF of each selected star is projected over the correlative part of the CCD. The new PSF is built by interpolation between PSFs corresponding to the most close values (aT and position) amongst the parametrized PSF models. This operations ends on a PSF image rebined and placed in the selected area of the CCD. The same operation is repeated for all the neighbours of the target included into the selected area and, at last, the background straylight level is added. Consequently the final image is composed of two parts: source and background, in order to determine the initial estimate of the S/N in each pixel. Absolute scaling is done in order to convert stellar magnitudes to photoelectron events. Such images are the basis to determine the patterns and, going ahead, to analyze the local photometric environnement of a particular target.

Full images are built in the same way adding targeted stars, background stars, photon noise, smearing, cosmic rays, star saturations, etc.. Main application of full images is to prepare the check, once COROT in orbit, that the targeted stars fit their forecasted parameters. For this a unique full CCD image will be downloaded at the begining of each observational period.

[†] we discriminate aCI from the true color index corrected from stellar extinction; to deduce spectral types and classes from color index is not a straightforward task because reddening in this regions is very frequent and it has to be accurately deduced aside

[‡] In order to set up an appropiate link between color index and the PSF spectral entries an ad-hoc parameter the "apparent temperature" (aT) is used

4. OPTIMAL MASKS

4.1. Build up process

In each field of view COROT will measure the set of target stars. CCD positions of these stars will be registered in the on board extractor. It will select for each position a specific pattern surrounding the star image. The number of admissible patterns is limited to 256 and they must be defined on ground. From the ground COROT receive 1) the set of patterns and 2) the set of positions, each with its reference to the assigned pattern.

As said in section 2, pattern definition is a multistep process using actual star fields and instrumental characteristics to determine the photometric environemment of each target in the CCD image. First step determines an optimal window for each of the $6000 \times n$ targets (where $n \sim 5$ is the number of fiels to observe) independently from the others. Second step reduce the collection to about 2000 sub-optimal masks grouping the star parameters. Third step reduces further the set sub-optimal masks to se the final set of 256 generic masks by similitude criteria. Fourth step assigns the best generic mask to each star.

The optimal window is deduced setting to zero the jitter and breathing and selecting the pixels such if added to measure they increase the S/N. Only the Poisson noise generated by the the targeted PSF itself, the neighbours and the background sources matter. To ensure the *only a segment per line and per target* condition the build up of each pattern is achieved line by line. Each target is synthetized from both: its own PSF image and the background image. Procedure follows:

1. The pixel with the better S/N and its line is chosen as seed. Lets be I the number of pixels in each line, (i,j)[§]. the pixel coordinates and $S(i,j), N(i,j)$ its respective signal and variances. Lets be (i_0, j_0) the "best S/N pixel".

2. For the line j_0 the S/Ns $r(.)$:
$$r(k, i_0; j_0) = \frac{S_O + \sum_{i=k}^{i=i0} S(i, j_0)}{\sqrt{V_0 + \sum_{i=k}^{i=i0} V(i, j_0)}}$$
are calculated for each point of the sequence $\{0, 1, \ldots, i_0 - 1\}$ (left sequence) and of the sequence $\{i_0 + 1, \ldots, I - 1\}$ (right sequence). The highest $r(k, i_0; j_0)$ defines the best stops points k^-_{j0} , k^+_{j0} for the left and for the right sequence respectively. S_0 and V_0 are the cumulated values of the signal and variance respectively.

3. The procedure defined in 2) is used in successive lines, $j_0 + 1, j_0 - 1, j_0 + 2, j0 - 2, \ldots$, starting with the "best" pixel on the line. S_0 and V_0 cumulate the S and V of accepted pixels in the successive lines, $S_0 = 0$ and $V_0 = 0$ for the initial line $j = j_0$.

4. Procedure stops when no pixel adjacent to previous segments exist.

The set of pixels defined by the previous process complies by construction with continuity requirements and is very close to the optimal mask. We call 'n theory" optimal mask the set defined: 1) sorting the pixels by S/N; 2) adding the flux and variance of pixels by decreasing order and 3) selecting the index where the S/N is the maximum. Such mask can not be used because it may be not compliant with the *only a segment per line and per target* requirement.

4.2. Contamination

In dense fields, a star PSF is often distorded by the close neighbours. This perturbing effet is called *contamination*. Contamination can modify the expected pattern mask dramaticaly. Contamination is measured here by the ratio between the flux due to the star neighbours in the measured flux and the flux due to the star itself. Very often star neighbours can include many referenced stars whose contribution exceeds 1% of the measured star. Of course bright stars show tipicaly low contamination $C < 0.01$ and dim stars show moderate $0.01 < C < 0.1$ to high contamination $0.1 < C$. The total flux, the global S/N, the contamination, the mean background level, the position and the color index are the key parameters defined by the first step.

[§]in conventional image coordinates i stands for column index and j for line index; $(0,0)$ is the left-bottom corner

4.3. Jitter and breathing disturbances

The Corot platform is 3 axis stabilized. Star trackers will maintain the platform in the right attitude for the preliminary phases and hand over this task to the sysmographic channel of COROT for the observational phase. Residual mouvement depend on the residual pointing error and on the weak oscillatig mouvement centered on the sysmographic star. Because the photometric mask have sharp borders and the background is variable around, any mouvement will modify the photometry. Therefore the *jitter* i.e. the pointing error vs. time $\lambda(t)$ will cause additional noise in the photometric measures.

A study, performed in 1D for sake of clarity, allows to quantify and reduce the jitter's noise, relying on different jitter models. In this study, X and Y are mask borders, $g(x)$ is the contaminated PSF value. Three scenarii being considered here:

1. *Random jitter:* Because the sampling is slow, this jitter is assimilated to white noise in depointing value. The additionnal term to total variance is:

$$V_J = \sigma_\lambda^2 \left(g(X) - g(Y)\right)^2$$

   This shows the interest of choosing mask limits close to equilevel lines of PSFs.

2. *Perfectly known jitter:* This ideal model attempts to reduce jitter effect by using additionnal information on jitter: it's time law. This law can be deduced from the serie of the sucessive on-board depointing measurements. Corrective actions are supposed to be intantaneous regarding the time between two of them. Then $\lambda(t)$ can be rebuilt, considering that between 2 actions, the satellite slides balistically from one known position to the next. In this case the additive term to total variance is null. Meanwhile, due to depointing, the S/N is reduced by two effets: 1) a little decrease of informative photons (i.e coming from the measued star) is lost out and 2) a corresponding increase of contaminating photons.

3. *Imperfectly known jitter:* In this actual model, $\lambda(t)$ is known with a samall error δ. The resulting added noise term is:

$$V_J = \delta^2 \left(g(X) - g(Y)\right)^2$$

   In this case, V_J is improved, compared to the random jitter model, in proportion of $\delta^2/\sigma_\lambda^2$. Moreover, the optimising citeria is the same in the two cases.

Due to the thermoelastic sensivity of the COROT telescope the night-day periods in the orbital trajectory will cause little but significant changes on the PSFs. This effect is known as *breathing*. For sake of tractability breathing is assimiled to a "extend/shrink" (c_{ES}) parameter neglecting other second order phenomena (lateral shift is included into the jitter variance).

$$V_{ES} = c_{ES} V_a$$

Similary to the *Known jitter* models c_{ES} depends on the accuracy in the breathing value restitution. Breathing is a very low frequency phenomena unlike jitter. Breathing will be detected using many parallel records. It is not possible at present to know how precisely can be measured and removed.

Jitter and breathing introduce new terms in the optimization procedure for patterns. However a theoretical move is excluded due to the overall complexity. In practice the starting point is the pattern defined in the optimal window procedure 4.1. Then we continue with the quasi-exaustive search for a minimum variance including jitter and breathing using the studied models.

In order to reduce the exaustive search initial masks are enlarged only by a pixel wide in four successive directions: $x+$,$x-$,$y+$ and $y-$. For each direction all admissible enlargements are tried (see fig.)

5. GENERIC MASKS

The restrictions of the on board extractor compel to replace the full set of initial masks with a reduced set of 256 generic masks. As said before this reduction is done in two steps first to 2000 masks then to 256 in a further reduction.

The first reduction is guided by decisions done in the space of star parameters, i.e. brightness, position and pseudo-temperature. In this space masks change smoothly. Contamination adds a new coordinate in the parameter space. Heavly contaminated stars show very specific patterns depending on its neighbours. Moderately contaminated stars generate masks with a "standard look". To deduce generic masks, specific windows are classed and added on large volumes of the parameter space, composing a reduced set of classes.

Parameter	N of Cls	Classes	Comments
mag(R)	7	$\{12., 12.5, 13., \ldots 15.\}$	Red magnitude
log(aT)	4	{3.863, 3.763, 3.707, 3.584}	log of apparent temp.
XC	3	$\{0^0.20, 0^0.86, 1^0.52\}$	angle of sight (X)
YC	6	$\{-0^0.320, 0^0.208, 0^0.736, 1^0.264, 1^0.792, 2^0.320\}$	angle of sight (Y)
C	4	{0.01, 0.1, 1.0, 10.0}	contamination level

The stars wich contamination exceeds 1. are discarded so this parametric space is divided in 1512 classes. Figure 1 shows a subset of 432 low contamined classes among the 1512. Each of the 6×3 cluster represents a value of YC $\times$ XC angle. In a given cluster, columns are log(aT) and rows are mag(R). A class mask is obtained by simply adding all specific masks class stars, followed by a thresholding. The gray level indicates the number of specific masks implied in each pixel.

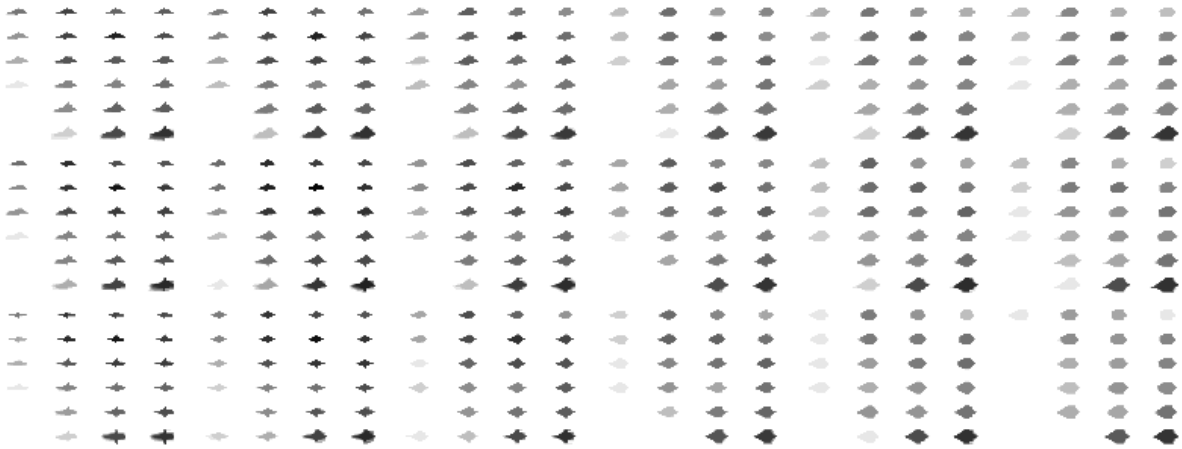


Figure 1. 432 Classe's Masks (see details in the text)

All the classes are grouped further in a second step in order to meet the requirement of 256 different classes. The classical PCA (principal component analysis) is used in this step.

6. EXPECTED PERFORMANCES

In the training work a limited field has been used for sake of performace. It was extracted from the DENIS and from the USNO catalogs and it covers a near square field of $0^0.6 \times 0^0.64$ The central sight point is $[284^0.29, -9^0.575]$. We found 17.648 stars of mag(R) in this field, where 1165 were selected as potential targets.

Figure 2 shows the number of concerned stars vs. $log_{10}(S/N)$. Graphics concern the 432 previous classes. One graphic per magnitude. There are 300 stars is the representative chosen field. The S/N is evaluated for a

single 32s exposure. Because each mesure sent by TM cumulate 16×32s integrated fluxes, its S/N is expected to be 4 times better $(log_{10}(S/N) + 0.6)$.

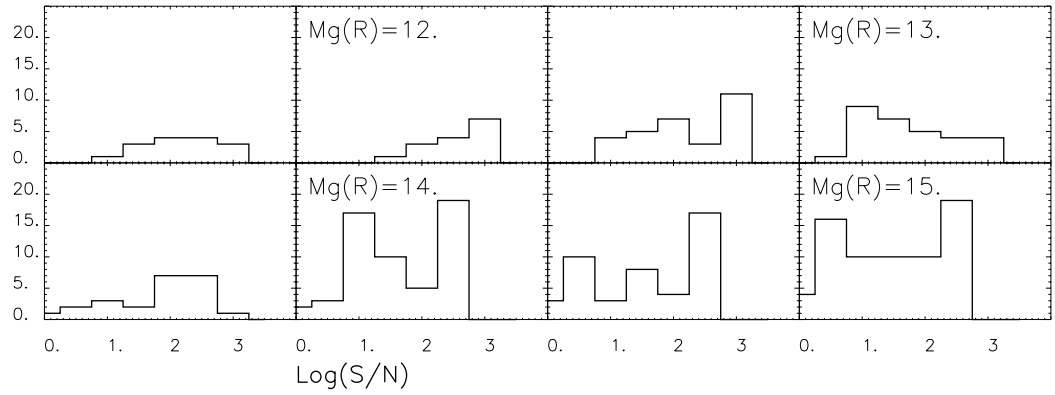


Figure 2. expected S/N(see details in the text)

7. CONCLUSION

The method described is able to determine a limited set of generic masks to monitor the star flux variability in a crowded field of COROT mission. The useful range spans star magnitudes (R) 12 to 14. We have here detailed the main steps leading to a collection of masks . The performances of the all system are being analysed and the preliminary studies show that the use of generic masks preserves for a the majority of cases the theoretical performances of the use of specific masks. Results are preliminary but show a clear trend and confirms the extreme importance to minimize jitter and breathing (J& B). A more complete study remains to be done using 1) the new set of PSFs 2) more actual J & B levels and 3) star variability (for 10% of stars).

APPENDIX A. FIELD SIMULATION

A.1. Characteristics of the star field

Due to the limits in catalog completude, $mag(R) = 15$ (approx.) is the frontier between two methods: for $mag(R) < 15$ only stars in catalogue are used for $mag(R) > 15$ star distribution is simulated using the local statistics The exact limit can vary depending on the maximal number of observed stars in the field. The upper limit (mag(R) $\sim$ 2021) depends on the background S/N.

A.2. Field simulation for stars of mag(R)< 15

Purpose: A precise prediction of each mini-spectra (for $\sim$ 20000 stars) and its surroundings Uses existing stars catalogues (DENIS, USNO). Each mini-spectra is placed with 1/10th of pixel position and rebined to CDD resolution. Each original mini-spectra (PSFs) depends on color index, position and magnitude. Algorithm :

$$I^{(n+1)} = E^{(n+1)}.PSF(c_{index}(n+1)) + I^{(n)}$$

where $E^{(n+1)}$ is the $n+1$ star and $I^(n)$ the simulated image in the n step

A.3. Field simulation for stars of mag(R)>15

Purpose: To predict the background of each mini-spectrum with statistical accuracy and fast enough($\sim 3.10^6$ stars) Catalogues are completed with random sets of fictive stars (distr. exp. for mag, uniform for a,d). Position accuracy: $\sim 1/4$ pixel . Background is built using a mean mini-spectrum (PSFm) Algorithm:

$$\begin{aligned} I_0^{(n+1)} &= E^{(n+1)} + I_0^{(n)} \\ I &= PSF_m * I_0^{(n+1)} \end{aligned}$$

APPENDIX B. JITTER STUDY

B.1. Random Jitter Model

Without jitter, the PSF function $G()$ ($, g()$ for contamined flux) spreads the flux $< F >$ ($, < f >$) into an $[X, Y]$ mask limits such as $< F >= \int_X^Y G(x)\,dx$. A λ depointing causes a λ-shift to mask limits. The resulting flux is:

$$< F|\lambda >= \int_X^Y G(x-\lambda) \tag{1}$$

In this part, we consider depointing as a probability distribution $P_\lambda[< \lambda >= 0, \sigma_\lambda^2]$. Lets compute S/N.

$$S =< F >= \sum_{m=0}^{\infty} mP(m)$$

Where $P(m)$ is the probability to receives exactly m photons during time exposure. Bayes theorem gives for signal:

$$\begin{aligned} S &= \sum_{m=0}^{\infty} m \int_\lambda P_{|\lambda}(m) P_\lambda(\lambda) = \int_\lambda < F|\lambda > P_\lambda(\lambda) \\ (1 \Rightarrow) &= \int_X^Y (\, G * P_\lambda\,)(x) \end{aligned}$$

For noise, assume $N^2 = \sigma^2 =< f^2 > - < f >^2$ of the jitted and contaminated flux. Since $f|\lambda$ is a Poisson variable,

$$\sigma^2 = \int_\lambda \left(< f|\lambda >^2 + < f|\lambda >\right) P_\lambda(\lambda) - \left(\int_\lambda < f|\lambda > P_\lambda(\lambda)\right)^2 \tag{2}$$

Thanks to $\lambda \ll Y - X$, we can develop at first order $< f|\lambda > (\lambda) = h + \lambda h'$. (2) becomes $\sigma^2 = h + \sigma_\lambda^2 h'^2$.

$$\text{knowing} \quad \left\{ \begin{array}{rcll} h &=& < f|\lambda > (0) &= \int_X^Y g(x) =< f > \\ h' &=& < f|\lambda >' (0) &= g(X) - g(Y) \end{array} \right.$$

We finally obtain $\boxed{S/N = \dfrac{\int_X^Y G * P_\lambda}{\sqrt{< f > + \sigma_\lambda^2 \,(g(X) - g(Y))^2}}}$

This shows that for a given jitter, noise increases with flux difference at PSF's borders. Hence, an optimal mask shall be close to an equilevel line of PSF.

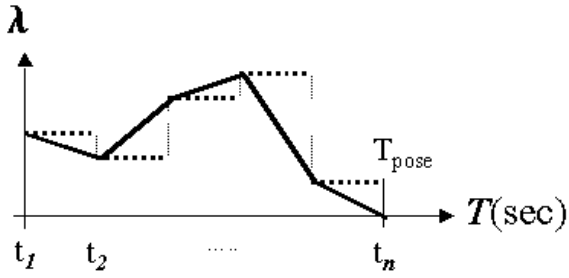


Figure 3. Jitter Balistique

B.2. Jitter Noise Reduction

The jitter noise can be reduced by the knowledge of depointing used for attitude corrections. Corrective actions $i, i+1$ are considered intantaneous regarding the time $t_i t_{i+1}$ between them. During $t_i t_{i+1}$ the satellite slides balistically from one position to the next. Then the jitter time law is known as serie of sucessive linear segments as shown fig 3

During time dt the depointing contributes for $< dN >=< F|\lambda > dt$ to total photons received. The measured flux during time exposure T is given by:

$$< F_{\text{mes}} >= \frac{1}{T} \int < dN >= \frac{1}{T} \int_0^T \int_X^Y G(x-\lambda) dx dt$$

Wich can be written as

$$< F_{\text{mes}} >=< F > -K \tag{3}$$

With K =constant=$\frac{1}{T} \int_0^T \left(\int_X^{X-\lambda} G(x)\, dx + \int_{Y-\lambda}^{Y} \cdots \right) dt$ (, k, g for contamined). K represents the part of star flux out of mask, due to jitter (, k is the part of additional contaminating flux in the mask).

The random variable $M = F_{\text{mes}} + K$ (, $m = f_{\text{mes}} + k$) is obviously an unbiased estimator of $< F >$(, $< f >$). Since var(m)=var(f_{mes})+var(k) =var(f_{mes}), we finally obtain:

$$\boxed{S/N = \frac{<F>-K}{\sqrt{<f>-k}}}$$

We recognize in N a photonoise of unjitted flux $< f >$, simply lacking it's portion k fell out of the mask, and added with extra contaminating photons.

Conclusion: Jitter's is *in theory* cancelled, thanks to signal m. m also permits to compute an *intantaneous* S/N for later processings. The limits are 1) accuracy in the depointing information, 2) accuracy of the PSF simulation. The first of this limits can be improved by using depointing history, or correlation across a large number of stars signal variations, pondered according to their positions on CCD.

B.3. Influence of Measure Accuracy

The depointing position is considered known with a certain error $-\delta$. Let's study δ's influence on noise reduction. Substituting $\lambda(t)$ by $\lambda(t)-\delta$, the expression of K (equat. 3) becomes:

$$K' = \frac{1}{T}\int_0^T \left(\int_X^{X-\lambda+\delta} G(x) + \int_{Y-\lambda+\delta}^{Y} G(x) \right) = K + \frac{1}{T}\int_0^T \left(\int_0^{\delta} G(x+X-\lambda) - \int_0^{\delta} G(x+Y-\lambda) \right)$$

Since $\delta \ll \lambda$, we can develop at first order, using $\int_0^{\delta} G \simeq 0 + \delta G(X-\lambda)$ (resp, $\delta G(Y-\lambda)$):

$$K' = K + \frac{\delta}{T}\int_0^T [G(X-\lambda) - G(Y-\lambda)]$$

This expression still depends on λ. Since $\lambda \ll X-Y$, we can develop again, using $G(X-\lambda) \simeq G(X) - \lambda G'(X)$ (resp, $G(Y) - \lambda G'(Y)$):

$$\begin{array}{rcll} K' & = & K + \delta[G(X) - G(Y)] + \delta\overline{\lambda}[G'(Y) - G'(X)] & \\ & \approx & \delta[G(X) - G(Y)] & \text{because} \overline{\lambda} \approx 0 \end{array}$$

Considering contamined fluxes, the noise added by residual jitter, compared to the term $V_J = \sigma_\lambda^2 \left(g(X) - g(Y)\right)^2$ obtained in §B.1 is improved in proportion of $\frac{\delta^2}{\sigma_\lambda^2}$.

APPENDIX C. BREATHING NOISE REDUCTION

Disfocusing the CCD, Breathing can be modeled by a PSF-centered dilatation. If M is the center of PSF, μ the dilatation coefficient, the measured flux becomes:

$$< F|\mu >= \int_X^Y \frac{1}{\mu} G(M + \frac{x-M}{\mu})\, d\mu = \int_{M-\frac{\overline{XM}}{\mu}}^{M+\frac{\overline{MY}}{\mu}} G(u)\, du$$

Wich can be decomposed in:

$$< F_{\text{mes}} >=< F|\mu >=< F > +L$$

With L =constant= $\int_{M-\frac{\overline{XM}}{\mu}}^{X} G(u) + \int_Y^{M+\frac{\overline{XM}}{\mu}} G(u)$ Similary to §B.2, we take advantage to use the unbiased signal $n =< F_{\text{mes}} > -L$ whose S/N is again:

$$\boxed{S/N = \frac{<F>-L}{\sqrt{<f>-l}}}$$

We again recognize a breathingless photonic S/N for a flux simply relieved from the out masked photons and added with extra enterring contamination.

C.1. Breathing With Jitter Noise Reduction

By combining §B.2 and §C the measured flux becomes

$$\begin{array}{rcl} < F_{\text{mes}} > & = & \frac{1}{T}\int_0^T \int_{X-\lambda}^{Y-\lambda} G_{\text{dilated}}(x) dx dt \\ & = & \frac{1}{T}\int_0^T \int_{M-\frac{\overline{XM}}{\mu}-\frac{\lambda}{\mu}}^{M+\frac{\overline{MY}}{\mu}-\frac{\lambda}{\mu}} G(u)\, du \\ & = & < F > +Q \end{array}$$

With Q =constant= $L + \frac{1}{T}\int_0^T \left[\int_{M-\frac{\overline{XM}}{\mu}-\frac{\lambda}{\mu}}^{M-\frac{\overline{XM}}{\mu}} + \int_{M+\frac{\overline{MY}}{\mu}}^{M+\frac{\overline{MY}}{\mu}-\frac{\lambda}{\mu}}\right]$
Using unbiased signal $p = < F_{\text{mes}} > -Q$ gives a S/N:

$$\boxed{S/N = \frac{<F>-Q}{\sqrt{<f>-q}}}$$

In wich we again recognize S/N of pure photonic noise for a flux simply relieved from the out masked photons and added with extra entering contamination.

C.2. Acknowledgments

We are grateful to E.Brier for his helpful discussions about the jitter statistics as well to M.Auvergne and M.Olivier for the fruitfull exchanges on mask determination and noise level determination. This work has been found by the CNES (the french space agency) and the CNRS.

REFERENCES

1. D. Rouan, A. Baglin, P. Barge, E. Copet, M. Deleuil, A. Léger, J. Schneider, D. Toublanc and A. Vuillemin "Searching for exosolar planets with the COROT space mission", *Physics and Chemistry of the Earth Part C*, **24, 5**, pp. 567–571, 2000.
2. D. Rouan, A. Baglin, E. Copet, J. Schneider, P. Barge, M. Deleuil, A. Vuillemin and A. Léger, "The Exosolar Planets Program of the COROT satellite", *Earth, Moon, and Planets*, **81, 1**, pp. 79–82, 2000.

Chapitre 6

Réduction Optimisée du nombre de patrons

6.1 Introduction au problème de la réduction optimisée

Le problème est le suivant : disposant de 12 000 étoiles toutes différentes par la combinaison de leur magnitude, spectre, géométrie de contamination, on doit réduire d'un facteur 50 la diversité de leurs masques optimaux tout en préservant la qualité des courbes de lumière. Ces deux nécessités sont contraires. En effet les étoiles ne pouvant conserver leur masque le mieux adapté, le S/B est forcément dégradé. L'objectif est donc de gérer au mieux cette contradiction.

N.B : Dans la pratique la réduction ne s'opérera pas forcément sur le champ à observer lui-même ni même sur un champ réel, mais il est commode de le considérer comme tel durant la description des méthodes.

La première des méthodes est une approche paramétrique qualitative qui consiste à regrouper les étoiles par familles partageant les mêmes critères *a priori*. Il s'agit de critères physiques traditionnels ; toutes les étoiles de la même famille adoptent un patron commun.

J'ai envisagé quatre autres méthodes de réduction qui sont exposées ci-après. Certaines sont des adaptations, d'autres sont conçues spécifiquement et chacune s'efforce de pallier les défauts de la précédente. Dans la deuxième technique, la classification *a posteriori*, je formalise les concepts de la méthode a priori pour appliquer la même opération après avoir identifié les facteurs effectivement dominants. Mais cette approche s'avère biaisée.

Le contrôle du biais m'a conduit à adopter une approche morphologique du problème. La troisième méthode est une condensation itérative où à chaque étape les deux masques les plus ressemblants fusionnent en un masque unique.

La quatrième méthode laisse les masques initiaux en place et répartit parmi eux 250 "graines" initiales (des masques au hasard) qui évoluent itérativement de sorte à minimiser la distance totale aux autres masques. Après convergence, les graines deviennent les patrons recherchés. Mais on constate qu'une apparence similaire entre masques et patrons n'est pas garante d'un bon S/B .

Puisqu'une performance accrue requiert un contrôle permanent du S/B , j'ai conçu la cinquième méthode, un tri efficace de masques aléatoires. C'est l'approche retenue pour Corot . Elle impose d'adopter dès le début une définition claire de la qualité globale d'un champ et de l'exprimer à l'aide d'un paramètre unique. Cette définition permet de jeter un nécessaire pont entre le S/B global et les S/B individuels. J'utilise ce paramètre comme métrique pour guider une sélection parmi un grand nombre de masques pseudo-aléatoires, après avoir quantifié la tolérance des étoiles envers des masques qui ne leur sont pas adaptés.

Ces différentes méthodes sont publiées dans l'article (Llebaria et al. 2003) inséré à la fin de ce chapitre.

6.2 Nécessité et contraintes de la réduction

Après chaque pose, les pixels sont lus à raison de 4Mbit/s. A une telle cadence, pour séparer les pixels à conserver des autres pixels on a recours à une électronique rapide de pré-traitement de type logique câblée, comme l'explique la note de Steller et al. (2002). Classiquement la distinction se fait par un bit à 1 ou à 0, selon que le pixel est à conserver ou non, dans une table en mémoire, miroir du CCD. Mais la quantité de mémoire requise pour les 4 millions de pixels que compte un CCD est excessive. La cartographie des fenêtres est donc codée de manière plus compacte sous forme de descripteurs ligne à ligne. Pour conserver une taille raisonnable de ces descripteurs, l'index pointant sur le patron est un mot de 8 bits, ne pouvant prendre que 256 valeurs distinctes. A l'issue, les pixels conservés sont suffisamment peu nombreux pour que le microprocesseur, plus lent, puisse se charger de les trier.

Parmi les 256 patrons disponibles, six ont des usages réservés, tels les fenêtres de mesure du fond, si bien qu'il n'en reste que 250 pour les étoiles cibles.

6.3 Méthode 1 : Paramétrisation *a priori*

Dans cette approche, on suppose qu'un patron qui donne un S/B élevé pour son étoile aura aussi un bon S/B sur d'autres étoiles similaires. La méthode (cf. Llebaria et al. (2002)) consiste à regrouper les étoiles par familles qui partagent des caractéristiques communes. Puis chaque famille reçoit un patron qui lui est propre. On choisit les *facteurs d'influence* qui président à la composition des familles parmi les *paramètres physiques* habituels en astronomie (les termes en *italique* seront repris dans la deuxième méthode). Des classes d'équivalence sont établies pour les facteurs suivants :

1. La magnitude des étoiles contraint l'aire de la PSF, elle est divisée en 7 plages.
2. La température de surface s'échelonne de 3500K à 9000K pour les cibles Corot . Elle contraint le spectre et donc l'amplitude des PSFs monochromatiques. On la sépare en 4 intervalles.
3. La position sur le CCD contraint la forme des PSFs. Elle est cloisonnée en 18 cadrans différents sur les deux CCDs.
4. La contamination joue sur la taille et la spécificité du masque optimal. Plus une étoile est contaminée, plus grand sera le nombre de pixels perdus à la frontière. On distingue 4 taux de contamination.

On obtient ainsi un millier de classes intermédiaires qui sont à leur tour regroupées pour passer sous la barre des 250 familles. Le patron commun à tous les membres d'une même famille est obtenu par moyenne des masques optimaux de la famille. Ses pixels n'étant plus binaires, sont *arrondis* à l'entier 0 ou 1 le plus proche. La figure 6.1 donne l'allure des masques réduits pour 432 classes d'étoiles peu contaminées.

L'approche a priori que nous venons de décrire donne un bon ordre d'idées mais reste assez qualitative et très arbitraire. En effet :

- Rien ne garantit que les paramètres habituels soient effectivement dominants pour Corot ;
- les frontières des classes sont fixées de manière arbitraire ;
- rien ne nous aide à pondérer leur importance ;
- il peut exister des paramètres cachés ou combinés propres à Corot ;
- les particularités statistiques du champ d'étoiles ne sont pas prises en compte ;

Malgré ses imperfections, cette méthode fournit une première approche du problème qui s'est avérée riche d'enseignements. Elle a permi de montrer que

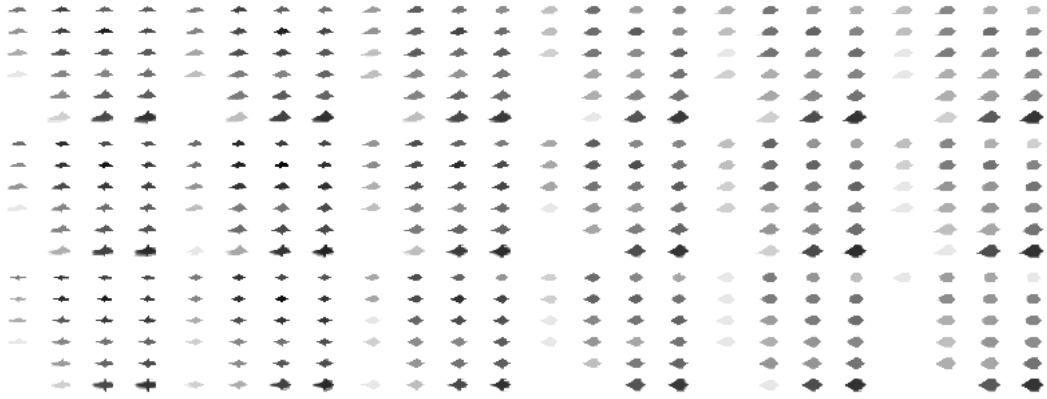


FIG. 6.1 – *Masques réduits pour 432 familles peu contaminées. Chaque bloc est un secteur du CCD. La température de surface varie suivant les colonnes et la luminosité suivant les lignes. Le niveau de gris indique le nombre de membres de la famille*

l'idée d'une réduction était réaliste sans hypothéquer irrémédiablement la précision photométrique. Nous allons au cours de la méthode qui suit conserver cette idée d'un regroupement par familles tout en la formalisant pour tenter d'en corriger les défauts.

6.4 Méthode 2 : Paramétrisation *a posteriori*

Nous reprenons la même approche de classement par famille, mais cette fois pour réduire le nombre d'hypothèses nous travaillons non pas sur les paramètres stellaires mais sur ce qu'en perçoit Corot c'est-à-dire les masques optimaux découlant des PSF sur le CCD. Nous ne cherchons plus à remonter aux paramètres stellaires.

6.4.1 Pertinence de la réduction :

Avant toute chose, commençons par vérifier qu'une réduction est bien envisageable. Pour cela, nous nous assurerons que les étoiles tolèrent d'autres masque que leur propre masque optimal. J'ai ainsi calculé le S/B sur 1 000 étoiles d'un échantillon témoin, lorsqu'on applique à chacune les 999 masques optimisés des autres étoiles. Le résultat est présenté sous forme de matrice figure 6.2. La dominance de stries verticales (à étoile constante) indique une préservation de l'ordre de grandeur du S/B ce qui légitime les tentatives de réduction.

Le long de cette description, nous expliciterons en **gras** les notions notées *en italiques* dans la première méthode. Afin de pouvoir comparer les différentes

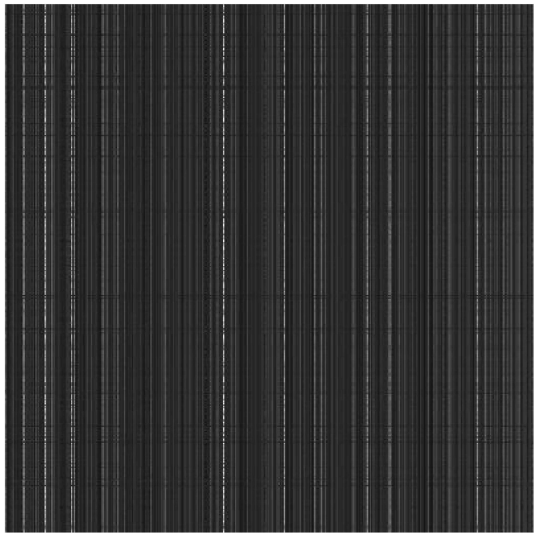


FIG. 6.2 – *Dans cette matrice* $1\,000 \times 1\,000$*, le point* q_{ij} *représente le S/B de l'étoile* j *(colonnes) à laquelle on applique le masque de l'étoile* i *(lignes). Un pixel clair indique un fort S/B . La dominance de stries verticales montre que la perte reste limitée en cas de masque désadapté. Des lignes sombres horizontales pointent les masques trop spécialisés qui ne s'adaptent pas à d'autres étoiles. La diagonale* q_{ii} *contient les masques optimisés. Elle n'apparaît pas assez contrastée dans cette représentation non-logarithmique. En effet la variation de S/B pour une étoile est faible devant les variations entre étoiles.*

méthodes nous garderons le même échantillon d'étoiles que celui que nous venons d'utiliser.

6.4.2 Dimension sous-jacente

Un masque est symbolisé par un vecteur $\overrightarrow{m}$ dont les coordonnées sont binaires (cf. Fig.6.3) Il possède une coordonnée par pixel, obtenue en mettant bout à bout les lignes des imagettes de travail 37×16. Une coordonnée égale à 1 signifie que le pixel de l'imagette est couvert par le masque. Initialement, $\overrightarrow{m}$ possède $n = 592$ coordonnées dans cette base, dite canonique.

On suppose que les patrons utilisent les mêmes pixels que les masques optimaux ce qui nous conduit à travailler dans le sous-ensemble $\mathcal{E}$ engendré par ces derniers. La vraie dimension de $\mathcal{E}$ est certainement inférieure à n. Pour la connaître on extrait la plus grande famille libre en réduisant la matrice $592 \times 1\,000$ formée des vecteurs colonne $\overrightarrow{m_i}$. On obtient moins de 100 coordonnées suf-

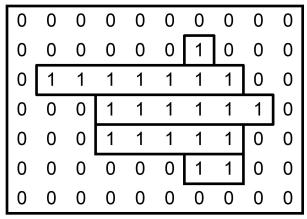


FIG. 6.3 – *Notation vectorielle des masques : la succession des lignes binaires forme un vecteur.*

fisantes pour décrire $\mathcal{E}$. Donc plus de 492 demeurent fixes (liés à la taille de la trame de travail) ou varient ensemble (liés par une cause physique sous-jacente). On effectue le changement de base pour travailler plus simplement. En contrepartie, le lien avec les masques est moins direct et les nouvelles coordonnées ne sont plus binaires mais ceci n'est pas gênant à ce stade.

6.4.3 Formalisation de la méthode utilisée précédemment

Les "**paramètres physiques**" a posteriori sont les combinaisons de coordonnées (*i.e* d'autres coordonnées) dont les variations expérimentales sont indépendantes les unes des autres. On identifie ces paramètres en procédant à une analyse en composantes principales (PCA).

A partir de l'ensemble des m_i masques, construisons la matrice de covariance intra-masque centrée :

$$G = E\left(\langle \overrightarrow{m_i}.\overrightarrow{m_i}^t \rangle\right) \tag{6.1}$$

. Ses éléments constitutifs sont les variances et covariances entre coordonnées :

$$G = \begin{pmatrix} \sigma_1^2 & \mathrm{cov}(1,2) & \cdots & \mathrm{cov}(1,n') \\ \mathrm{cov}(2,1) & \sigma_2^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ \vdots & \vdots & & \sigma_{n'}^2 \end{pmatrix}$$

L'examen de G montre beaucoup de termes non nuls hors diagonale. Donc les coordonnées utilisées sont interdépendantes. G est symétrique par construction mais aussi définie et positive en tant que somme de carrés. Elle est donc diagonalisable avec des valeurs propres réelles et non négatives et des vecteurs propres qui forment une base orthogonale. Sa diagonalisation nous fournit les matrices D et P toutes deux de dimension $n' \times n'$ et telles que :

$$G = P^{-1}.D.P$$

D est la matrice diagonale des valeurs propres λ_i et P est la matrice de passage vers la base des vecteurs propres $\left\{\overrightarrow{V_1}, \ldots \overrightarrow{V_{n'}}\right\}$. P est formée des $\left\{\overrightarrow{V_i}\right\}$ en colonnes, dans l'ordre des λ_i.

Changeons de base et travaillons dans la base des vecteurs propres. Un masque s'y écrit $\overrightarrow{m} = \sum x_i \overrightarrow{V_i}$. Dans cette base la nouvelle matrice de covariance est directement D. Les V_i varient indépendamment les uns des autres car les termes croisés $\mathrm{cov}(x_i, x_j)_{i \neq j}$ sont tous nuls : Ce sont les **paramètres physiques** a posteriori recherchés.

Les "**facteurs d'influence**" sont parmi ces nouveaux paramètres, ceux qui entraînent le plus de diversité dans la forme des masques, c'est-à-dire les composantes qui varient le plus. Les autres peuvent être considérés comme une constante. L'"influence" d'un facteur V_i est donc mesurée par sa variance λ_i, les facteurs d'influence sont les V_i associés aux plus grandes valeurs de λ_i. En réordonnant les λ_i on obtient la répartition de la figure 6.4.

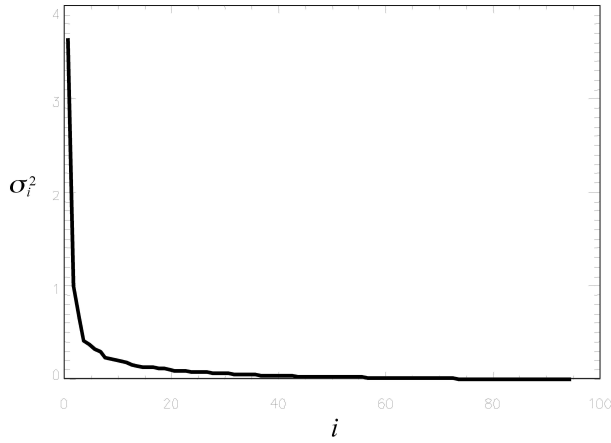


FIG. 6.4 – *Classement décroissant des valeurs propres. Le rang de la composante V_i figure en abscisse et sa variance en ordonnée. La dispersion σ_i^2 n'est importante que pour un petit nombre de paramètres. La variabilité totale est donnée par l'aire située sous la courbe.*

Un petit nombre de vecteurs propres, environs 5, suffit à exprimer la plus grande partie de variabilité des masques optimaux. Pour visualiser les V_i sous forme de facteurs d'influence, il faut revenir à la base canonique par inversion des changements de bases. La figure 6.5 présente un exemple de décomposition.

Une **famille** est par définition un groupe de masques dont les membres par-

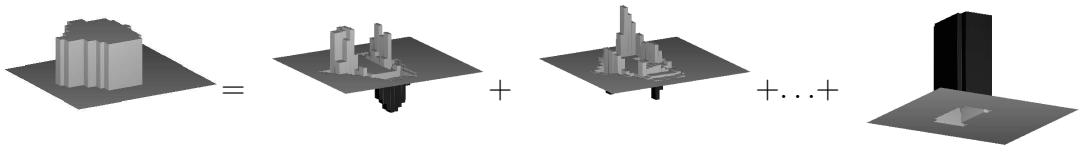


FIG. 6.5 – *Le masque m résulte de la somme des facteurs d'influence (ici non-pondérés), et de la constante. Les "pixels" utilisés à ce stade n'ont aucune raison d'être binaires, ni même positifs. Ce ne sont que des intermédiaires de calcul dont la somme* $\vec{m} = \sum x_i \vec{V_i}$*, elle, doit être binaire.*

tagent des caractéristiques proches, c'est-à-dire des coordonnées $x_1, \dots x_5$ voisines. Pour obtenir 250 familles, on partitionne les valeurs prises par x_1 (respectivement $x_2, \dots$) en k_1 (respectivement $k_2, \dots$) intervalles tels que :

$$k_1 \times k_2 \times k_3 \times k_4 \times k_5 \leqslant 250$$

Chaque combinaison $(k_1, k_2, k_3, k_4, k_5)$ est une famille a posteriori. Pour déterminer les limites définissant la famille i on examine la répartition des x_i. La figure 6.6 en montre un exemple. Une fois les familles cloisonnées, les masques optimaux de chacun de leurs membres sont moyennés pour fournir leur patron commun.

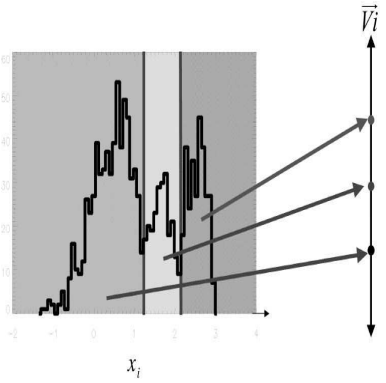
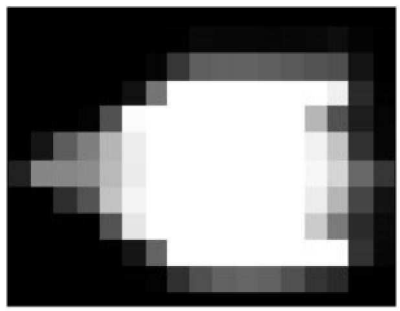


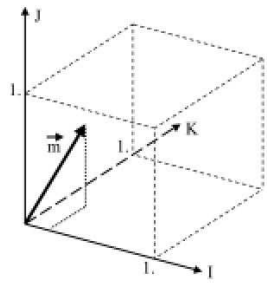
FIG. 6.6 – *Histogramme des coordonnées suivant un* $\vec{V_i}$*. Les frontières sont choisies dans les zones les moins peuplées.*

Arrondi : Ramené dans la base canonique, un patron $\overrightarrow{p}$ n'a aucune chance d'être binaire, comme dans l'exemple de gauche de la figure 6.7. Il faut donc identifier le patron binaire $\overrightarrow{p'}$ le plus proche. On remarque que les patrons binaires

sont les sommets de l'hypercube unitaire de dimension canonique 592. Le schéma de droite de la figure 6.7 montre que le problème se découple en projetant $\overrightarrow{p}$ indépendamment sur chaque axe. On choisit pour $\overrightarrow{p'}$ le coin le plus proche. Ce découplage justifie le **seuil** de $\frac{1}{2}$ appliqué lors de l'arrondi dans l'approche a priori.



Exemple de patron : ils ne sont plus binaires.



On remarque que leurs coordonnées continues $x_i \leqslant 1$ les rendent équivalents à un point à l'intérieur de l'hypercube unitaire de même dimension.

FIG. 6.7 – *Rebinarisation*

6.4.4 Défauts de cette approche

Malgré les améliorations apportées par cette méthode la détermination des frontières entre familles conserve une part d'arbitraire. Cette méthode est biaisée : si n =12 000 patrons étaient autorisés, elle ne convergerait jamais vers les n masques optimaux de départ. Nous ne poursuivons donc pas l'approche par familles et recherchons donc une méthode évitant ce biais.

6.5 Méthode 3 : L'homogénéisation morphologique directe

Afin de supprimer le biais, nous cherchons avec les méthodes 3 et 4 à maximiser la ressemblance morphologique entre masques et patrons. L'idée de la troisième méthode est de jouer à l'envers le film de la convergence : partant des 12 000 masques optimaux, nous nous en éloignons graduellement en réduisant

un à un leur nombre jusqu'aux 250 patrons recherchés. La quatrième méthode équirépartit directement les patrons parmi les masques. Après avoir exposé ces deux techniques, l'une basique et l'autre éprouvée, nous conclurons à la persistance de défauts.

Deux masques se "ressemblent" s'ils ont en commun un grand nombre de pixels. Les masques sont à présent symbolisés par des points dans la base canonique. La ressemblance de deux masques m_1 et m_2 est la distance quadratique :

$$\mathrm{D}(m_1, m_2) = \|\overrightarrow{m_1 m_2}\|^2$$

Etant quadratique, cette distance tend à éviter les écarts importants entre masques et patrons.

6.5.1 Algorithme de Base : la Condensation

Il n'existe pas de solution analytique connue fournissant directement le jeu de patrons qui minimise la distance totale des masques aux patrons (le grand minimum). En revanche on s'en approche en cheminant le long d'une suite d'étapes simples que l'on sait optimiser individuellement. Le minimum local qui est alors atteint n'a aucune raison d'être le grand minimum mais il peut nous suffire.

Nous allons jouer à l'envers le scénario de la convergence et supprimer les 12 000 masques optimaux un à un. Le mode de suppression est choisi pour minimiser à chaque fois le supplément de distance créée. A chaque itération un couple de masques-points est remplacé par un point unique à mi-chemin entre les deux. Pour prendre une analogie, ce processus dans son principe ressemble à la coalescence des gouttelettes d'eau lors de la condensation de vapeur sur une surface embuée : Quand deux gouttelettes entrent en contact, elle fusionnent en un goutte unique située quelque part entre les deux positions. Cette idée est à l'évidence sans biais.

Afin de déterminer la meilleure substitution nous cherchons a la première itération par quel point g remplacer un couple arbitraire $\{m_1, m_2\}$ pour minimiser l'accroissement de distance δ.

$$\delta = \overrightarrow{gm_1}^2 + \overrightarrow{gm_2}^2 \tag{6.2}$$

Par dérivation, on trouve que la solution g est le barycentre de $\{m_1, m_2\}$.

$$\frac{\partial \delta}{\partial \overrightarrow{og}} = 0 \Longleftrightarrow \overrightarrow{og} = \frac{\overrightarrow{om_1} + \overrightarrow{om_2}}{2} \tag{6.3}$$

Il reste à choisir le meilleur couple à substituer. D'après l'équation 6.2 l'accroissement de distance obtenu avec g vaut :

$$\delta = \frac{(\overrightarrow{om_2} - \overrightarrow{om_1})^2}{2}$$

Pour minimiser δ, il faut minimiser $\|\overrightarrow{m_1 m_2}\|$. La première itération consiste donc à choisir les deux points les plus proches et à les remplacer par leur barycentre. On affectera g du poids $w = 2$.

Généralisons ce résultat aux itérations suivantes qui agissent sur un "mélange" de points et de barycentres. Tout d'abord on constate que remplacer un couple de barycentres $\{g_1, g_2\}$ affecté des poids $\{w_1, w_2\}$ par leur barycentre g minimise effectivement la distance créée entre g et tous les points d'origine impliqués dans g_1 et g_2. En effet en généralisant l'équation 6.2 aux $w_1 + w_2$ points d'origine et en la dérivant on obtient :

$$\frac{\partial}{\partial \overrightarrow{og}} \sum_{w_1+w_2} (\overrightarrow{og} - \overrightarrow{om_i})^2 = 2 \sum_{w_1+w_2} \overrightarrow{og} - 2 \sum_{w_1} \overrightarrow{om_i} - 2 \sum_{w_2} \overrightarrow{om_j} \tag{6.4}$$

Le classement des points en m_i ou m_j est fonction de leur appartenance au barycentre g_1 ou g_2. On reconnaît dans les deux derniers termes la définition de $w_1 \overrightarrow{og_1}$ et $w_2 \overrightarrow{og_2}$. La dérivée s'annule bien pour :

$$\overrightarrow{og} = \frac{w_1 \overrightarrow{og_1} + w_2 \overrightarrow{og_2}}{w_1 + w_2}$$

L'accroissement de distance est bien minimal en remplaçant g_1 et g_2 par g. La première itération n'est que le cas particulier où $w_1 = w_2 = 1$, le terme général de l'itération sera obtenu en remplaçant les deux points les plus proches, barycentres ou non, par leur barycentre pondéré.

Bien que sans biais et apparemment logique, cette méthode est inefficace :

1. La quantité de calculs est excessive. Pour trouver les points les plus proches, il faut évaluer environ $n(n+1)/2$ distances à chacune des $\simeq n$ itérations, chaque distance nécessitant elle-même 592 multiplications. Cette complexité d'ordre n^3 n'est pas envisageable pour nos milliers d'étoiles. Il est cependant possible de la réduire en cloisonnant les points dans des boîtes virtuelles et en limitant le calcul aux distances inter et intra-boites. De plus, on peut travailler dans une base plus simple. La complexité

tomberait alors à n^2, valeur qui serait acceptable.

2. Le résultat est piètre dans la pratique. Tous les points condensent vers le plus petit masque. Ceci semble tenir au fait que, les petits masques ayant peu de coordonnées non nulles, sont plus proches entre eux que les grands. Ils commencent donc à se condenser, puis vident graduellement les couches périphériques, isolant un peu plus les grands masques. On peut y voir un défaut dans le choix de la distance qui n'est pas relative à la taille du masque. Mais il serait inutile d'adopter une distance normée par la surface :

$$\mathrm{D}(m_1, m_2) = \frac{\|\overrightarrow{m_1 m_2}\|^2}{\|\overrightarrow{m_1}\|^2 + \|\overrightarrow{m_2}\|^2}$$

Ceci aurait un effet désastreux : annuler l'importance de la surface du masque, et donc la brillance de l'étoile. Cette absence de succès met au grand jour la lacune évoquée au début de cette section : minimiser n accroissements locaux ne revient pas à minimiser l'accroissement global.

6.6 Méthode 4 : Le problème à K-moyennes

Nous allons voir que l'on résout cette difficulté en ne supprimant pas les masques optimaux initiaux, mais en les laissant coexister en permanence avec les patrons. Notre problème est analogue au problème connu dit *problème à K-moyennes*. Le terme en est : Etant donné n élèves répartis au hasard, où placer $p < n$ écoles pour minimiser le trajet total ? Il existe une classe d'algorithmes itératifs portant le même nom que le problème, décrite par Press et al. (1997). Ces algorithmes possèdent de "bonnes" propriétés : ils nécessitent peu de calculs et convergent rapidement vers une solution stable. Ils ont besoin d'une condition initiale arbitraire, mais le point de convergence est relativement indépendant de son choix. Dans notre cas $p = 250$ et n =12 000.

Le principe de l'algorithme à K-moyennes est très simple : On jette p "graines" au hasard (des masques qui vont devenir les futurs patrons) qui vont évoluer pour se répartir au milieu des masques. Contrairement à l'algorithme de condensation, celui-ci est protégé du risque d'accumulation car la position des masques optimaux initiaux reste inchangée. L'algorithme s'arrête quand un certain critère de convergence est atteint.

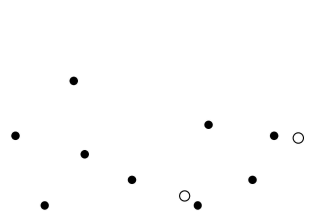
J'ai choisi la variante suivante :

Initialisation : Les 250 graines sont choisies au hasard parmi les 12 000 masques. C'est ce choix qui assure l'absence de biais : si 12 000 graines

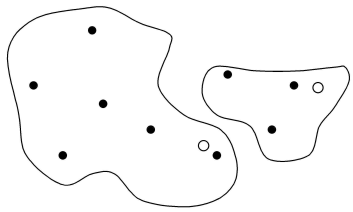
étaient admises, elle ne pourraient qu'être les 12 000 masques optimaux eux-mêmes.

Itération : La boucle itérative comporte deux étapes (voir Fig. 6.8) :

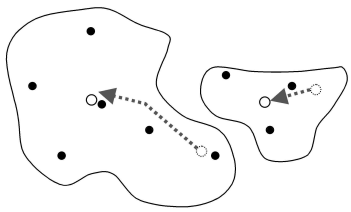
1. La formation des groupes. Il y a un groupe par graine constitué des masques optimaux les plus proches,
2. Le centrage de la graine. La graine est déplacée au centre de son groupe, c'est-à-dire le barycentre. Ainsi la distance totale entre graine et groupe est minimisée, rendant la graine représentative de son groupe.



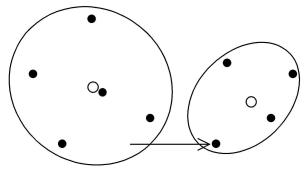
1) Les graines sont les cercles, les masques optimaux sont les points.



2) Formation des groupes : Chaque masque est affecté à sa graine la plus proche.



3) La graine est déplacée au barycentre de son groupe (la distance graine-groupe est minimisée).



4) Les nouveaux groupes sont formés etc ...

FIG. 6.8 – *Boucle d'itération.*

Arrêt : Au gré des itérations la distance totale diminue. La convergence absolue est atteinte quand les graines ne se déplacent plus. Mais on ne peut garantir l'absence d'oscillations infinies, ne serait-ce qu'à cause des erreurs d'arrondi. Nous avons amélioré ce critère par le choix d'une condition double : La convergence est atteinte si :

1. La distance totale ne diminue plus (moins de $1/1000^e$ deux fois de suite) ce qui englobe la convergence absolue,

ou

2. L'arrêt est inconditionnel si la convergence n'est pas atteinte après 20 itérations. Ce cas étant anormal, un message d'avertissement est émis.

Arrondi : comme précédemment les graines finales sont des moyennes et donc leurs coordonnées sont continues. On choisit pour chacune d'elles le masque binaire le plus proche (qui n'est pas nécessairement un masque optimal) qui devient un patron, comme on l'a vu au §6.4.3.

La complexité calculatoire est très réduite. Il ne faut pas plus de $n \times p$ évaluations de distance par itération, et dans notre cas la convergence est très rapide. Elle se produit la plupart du temps avant le 10^{e} tour.

Pour comprendre en quoi les graines ont tendance à se "repousser" plutôt qu'a "s'attirer", imaginons deux graines très proches à un moment donné. Elles laissent un grand trou Z_1 vide de graines. Ce trou n'en est pas moins peuplé de points qui deviennent autant d'attracteurs. Lors de la prochaine formation des groupes, ces points ne se trouveront rattachés qu'à la plus proche des deux graines. Celle-ci se déplacera alors vers le centre de Z_1 tendant à le combler, laissant elle-même un trou Z_2 plus petit derrière elle. A son tour Z_2 deviendra un attracteur, plus faible que Z_1, propageant ainsi un mouvement aux autres graines tout en l'atténuant. C'est pour cette raison que la distance finale est peu sensible aux conditions initiales.

Il existe de nombreuses variantes : La répartition aléatoire initiale peut être plus ou moins orientée ; un taux d'apprentissage peut freiner progressivement le déplacement des graines pour favoriser la convergence ; les points peuvent avoir des poids différents afin d'en privilégier certains ; enfin les graines peuvent être animées d'une faible "vibration" aléatoire additionnelle pour les faire ressortir des minimum locaux (dans ce cas la convergence devra tenir compte de cette vibration).

Résultat : sur notre échantillon de 1 000 masques à réduire en 40 patrons, les performances en termes de distance sont très bonnes. La convergence est obtenue en moins de 10 itérations et la distance totale est en moyenne inférieure à 2% de la taille du masque, soit 2 pixels pour les grand masques et 1 seul pour les petits.

Mais en dépit de ce résultat, cette méthode ainsi que toutes celles qui sont basées sur la maximisation d'une ressemblance sont inadéquates. Le résultat en termes de S/B est trop imprédictible. La figure 6.9 montre que malgré une distance égale entre un masque individuel et plusieurs patrons, le rapport signal à bruit varie beaucoup trop.

En fait, trop de facteurs entrent en jeu pour qu'il y ait continuité entre la forme d'un masque appliqué à une étoile et le S/B qui en résulte. Certains produisent des effets de seuil, tel le passage de coordonnée continue à coordonnée binaire, couplé avec la dépendance envers l'ensemble de la frontière. Pour compenser l'erreur due

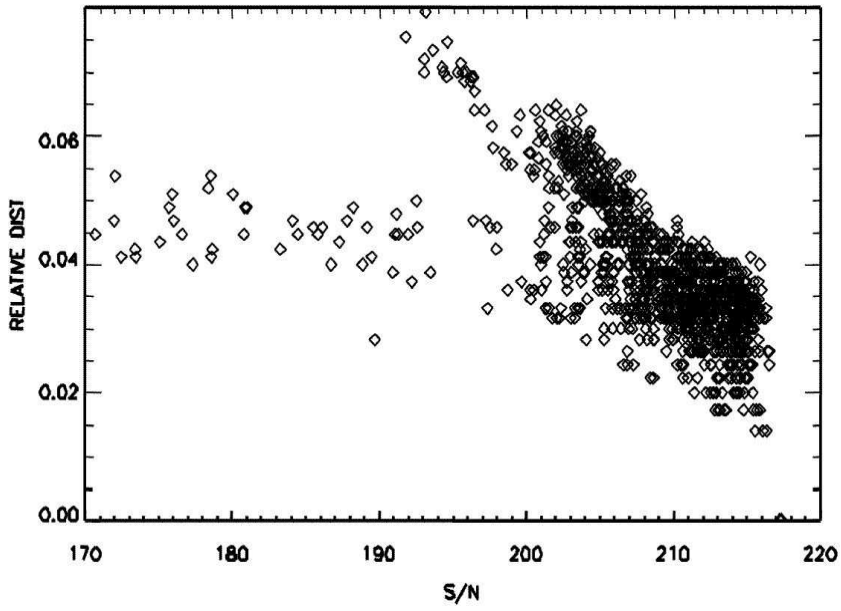


FIG. 6.9 – *Corrélation entre distance et S/B . Le S/B en abscisse est calculé pour des patrons situés à différentes distances (en ordonnée). La distance est relative au masque optimal (celui qui est situé à* D = 0*). L'étalement des points suivant les horizontales montre que pour une distance donnée le S/B est peu prédictible.*

à l'arrondi d'un pixel, il peut être nécessaire d'en faire basculer de nombreux autres. En conclusion le critère de ressemblance est peu légitime. Comme on souhaite une performance en termes de S/B , il faut trouver une métrique qui tienne compte explicitement du S/B .

6.7 Méthode 5 : Le tri efficace de masques pseudo-aléatoires

Nous décrivons ici la méthode utilisée en pratique, c'est la plus efficace de toutes car uniquement fondée sur le S/B .

Il faut tout d'abord fixer une définition du S/B global qui traduise au mieux les besoins scientifiques de Corot et qui s'exprime à l'aide d'un paramètre unique. Ce paramètre peut alors être utilisé comme métrique pour guider la réduction. On ne peut pas se contenter d'utiliser le S/B moyen d'un champ. En effet, une moyenne peu dégradée par la réduction peut néanmoins cacher des biais importants tels que la perte des meilleures cibles.

Le critère choisi consiste à limiter la dégradation de S/B que subit une étoile lorsque son masque optimal est remplacé par le patron générique qu'on lui attribue. Autrement dit, on s'impose qu'après réduction, toutes les étoiles considérées comme cibles conservent au moins

$$\mathrm{S/B} \geqslant \tau \times \mathrm{S/B}_{opt} \tag{6.5}$$

où $\mathrm{S/B}_{opt}$ est leur rapport signal à bruit que donne le masque optimal. Ainsi, la répartition finale des S/B est identique à la répartition initiale, les étoiles préservant leur capital dans une égale mesure.

Ce choix établit la correspondance entre S/B global d'un champ et S/B individuels des étoiles qu'il contient. En effet si pour toute étoile i l'on à :

$$\mathrm{S/B}'_i \geqslant \tau \mathrm{S/B}_i$$

où $\mathrm{S/B}_i$ est la valeur du S/B en amont de la réduction et $\mathrm{S/B}'_i$ sa valeur en aval, on sera alors assuré que

$$\overline{\mathrm{S/B}'} \geqslant \tau \overline{\mathrm{S/B}}$$

où $\overline{\mathrm{S/B}}$ et $\overline{\mathrm{S/B}'}$ sont les S/B moyens sur tout le champ avant et après réduction.

Afin d'arbitrer la contradiction entre le nombre réduit de patrons et la qualité du champ nous commençons par imposer une consigne τ, puis nous examinons le nombre de patrons requis $n = f(\tau)$ pour satisfaire cette consigne à l'aide d'un l'algorithme de tri f. L'algorithme de tri est décrit plus loin.

6.7.1 Dilemme de l'optimisation collective

Nous développons ici les raisons du choix du critère τ pour Corot .

Quand le nombre de patrons n'est pas limité, les optimisations locales et globales coïncident. L'amélioration du S/B d'une étoile profite au champ dans son ensemble. Mais quand le nombre de patrons disponibles diminue, les ressources disponibles pour récupérer l'information photométrique deviennent limitées. Ce qui est optimisé pour une étoile ne peut l'être pour les autres ; en quelque sorte l'intérêt du groupe n'est plus l'intérêt de ses membres. Le premier problème est donc de déterminer quelle répartition de S/B offrira le meilleur retour scientifique, avant même de chercher comment l'obtenir. Deux possibilités opposées s'offrent à nous :

– 1) Une répartition "inégalitaire" des ressources où les quelques étoiles les plus brillantes sont préservées au détriment de la majorité des autres, cela

sans égard pour le nombre de ces dernières,

– 2) Une répartition "égalitaire" qui équilibre l'usage de la ressource, et donc favorise les étoiles faibles en misant sur leur grand nombre.

On peut tenter d'échapper à ce choix de façon qualitative (voir Fig. 6.10 ci-dessous).

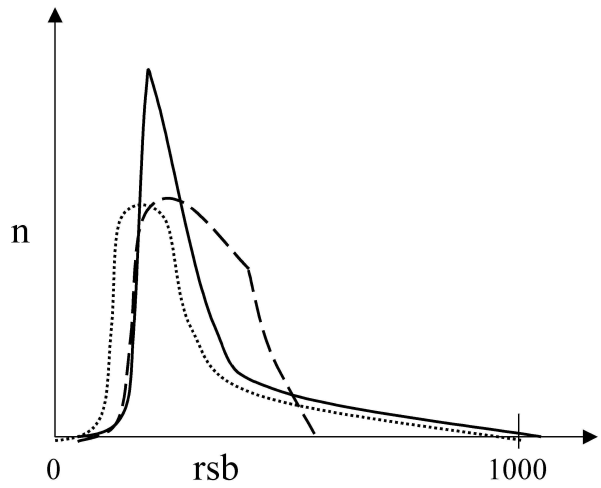


FIG. 6.10 – *Stratégies "égalitaires" et "inégalitaires". En trait plein le S/B hors contraintes : le patron de chaque étoile est son meilleur masque. En bistre la stratégie égalitaire : aucune étoile n'est privilégiée ce qui maximise le S/B moyen mais sacrifie les étoiles brillantes. En pointillés la politique inégalitaire où 20% des étoiles à plus fort S/B sont préservées au prix de 80% des patrons. Le S/B moyen n'est pas maximum.*

La performance de la stratégie utilisée peut être mesurée en calculant l'aire comprise sous les courbes de la figure 6.10.

$$\mathrm{Q} = \sum n_i.q_i$$

où n_i est le nombre d'étoiles conservant le S/B q_i après réduction. Ainsi on pourrait chercher à maximiser Q. Mais cette métrique cache en réalité la stratégie égalitaire en gommant le poids des disparités.

Une autre idée est l'introduction d'un facteur de "coût normalisé" C qui représente la "part" du jeu de patrons que requiert une étoile. C est élevé quand l'étoile est peu tolérante et nécessite un patron très spécifique, peu réutilisable par d'autres. Notre qualité devient alors :

$$\mathrm{Q} = \sum_i q_i.\mathrm{C}_i$$

Plus une étoile i "consomme" de patrons, plus elle doit être optimisée pour compenser son coût. Ces étoiles sont les étoiles brillantes ; le "coût" des étoiles faibles est réduit car leur signal photonique se distingue peu du bruit de fond (identique pour toutes les étoiles) et leurs masques tendent donc à se ressembler. La pondération utilisée traduit le fait qu'une stratégie inégalitaire n'est pas forcément pénalisante. Dans la pratique cette pondération ne sera pas utilisée. En effet, le facteur C est trop difficile à évaluer car il dépend de l'ensemble des patrons et des étoiles.

Nous recentrons maintenant ces développements en revenant aux besoins scientifiques de la mission Corot qui demande de conserver la meilleure précision sur les objets les plus brillants.

On utilisera donc la définition suivante pour la qualité d'un champ dont les masques optimaux ont été réduits : *la répartition des S/B dans un champ muni de patrons doit être la même que la répartition des S/B avec les masques optimaux.*

6.7.2 Mesure du S/B global

Une fois choisi le critère de qualité globale, il faut le mesurer à l'aide d'un paramètre unique. On pense en premier lieu à l'écart quadratique entre histogramme des S/B initiaux et finaux, mais une telle mesure n'a pas de sens individuel : les pertes de S/B risquent de se compenser entre étoiles brillantes et étoiles moyennes.

Pour qu'aucune étoile n'échappe à la règle il fallait un critère ferme, à base de seuil : la distance maximale entre ces histogrammes. C'est pourquoi nous avons choisi τ.

6.7.3 Tolérance aux masques aléatoires

Il est important de remarquer que le S/B est, en moyenne, assez bien préservé en valeur relative même si on mesure une étoile à l'aide d'un masque mal adapté. Cela apparaît dans la matrice étoiles $\times$ masques (Fig. 6.11), chaque point clair situé hors de la diagonale des masques adaptés atteste d'une bonne préservation du S/B . Ces points se rangent parfois le long de lignes horizontales claires, révélant que le masque est adapté à de nombreuses étoiles. Ce sont les masques correspondants à ces lignes claires qui deviendront les patrons. L'histogramme de la même figure montre que 45% des masques préservent en moyenne $\tau \geqslant 95\%$ de S/B aux étoiles. Les écarts peuvent être importants entre magnitudes, mais

cette valeur nous servira d'ordre de grandeur.

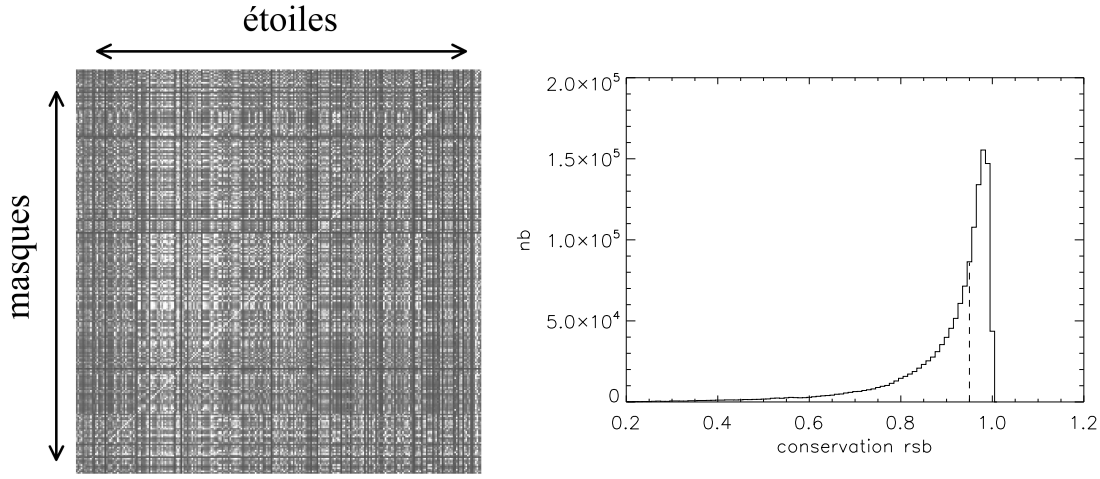


FIG. 6.11 – *a) Préservation du S/B avec des masques non adaptés. La matrice est extraite de la matrice étoiles/masques[1 000,1 000] de la figure 6.2 dont on à normé chaque colonne (étoile) par le S/B obtenu avec son masque adapté. On remarque les masques adaptés sur la diagonale. b) Histogramme des valeurs de la matrice de gauche. Jusqu'à* 45% *des couples masque* × *étoile (aire à droite de la ligne pointillée) perdent moins de* 5% *de S/B avec un masque désadapté.*

On va exploiter au mieux les recoupements au sein de ces 45% de masques. Le rôle de l'algorithme sera de rechercher les masques les plus communs possible parmi tous ceux qui préservent une étoile donnée.

Nous donnons maintenant une estimation qualitative de l'ordre de grandeur du nombre de patrons pour $\tau = 95\%$, sous l'hypothèse d'une répartition uniforme. Pour illustrer les choses supposons que les étoiles sont des boules stockées dans une urne et qu'un masque est un tirage des $p = 45\%$ de boules satisfaisant τ. Calculons après combien de tirages (avec remise) toutes les boules ont été tirées au moins une fois. La probabilité $\mathcal{P}(n)$ qu'une boule soit tirée au moins une fois en n tirages avec remise est :

$$\bar{\mathcal{P}} = (1-p)^n$$

Si on se donne moins d'une chance sur mille qu'une des boules de notre échantillon de 1 000 étoiles reste non-tirée, il faut $\bar{\mathcal{P}} \leqslant 10^{-6}$ par boule. La solution en n de l'inéquation est :

$$n \geqslant \log(10^{-6}) / \log(1-p)$$

Il faut $n \geqslant 24$ tirages (patrons). C'est bien dans la proportion recherchée car $24/1000 \simeq 250/12\ 000$, laissant espoir pour la réduction avec 95% de préservation.

6.7.4 Description de la méthode

La séquence complète de réduction est la suivante :

1. Initialisation : Le point de départ est un champ d'étoiles et un réservoir de masques aléatoires, la matrice des S/B s est obtenue par application de tous les masques à chacune des étoiles. Il est préférable que les masques aléatoires soient en fait les masques optimaux des autres étoiles du champ mais ce n'est pas indispensable.

2. Itération : Ces étapes sont illustrées figure 6.12.
   (a) On se donne une consigne sévère, par exemple $\tau = 0.97$,

   (b) La matrice des signaux à bruit est seuillée avec τ ce qui la transforme en matrice binaire, nommée matrice "d'acceptabilité" et dont l'élément $(i, j) = 1$ signifie "l'étoile j accepte le masque i". Il y a au minimum un masque acceptable par étoile : son masque optimal. Plus une colonne compte d'éléments égaux à 1, plus l'étoile est "tolérante". Plus une ligne compte d'éléments égaux à 1, plus le masque est "apprécié" des étoiles candidates.

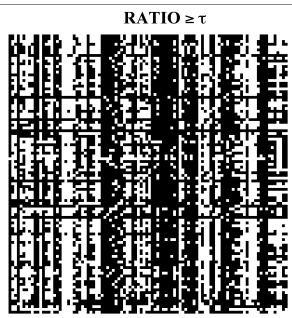
   (c) Le masque le plus apprécié de tous est choisi comme le premier des patrons,

   (d) Les étoiles qui ont accepté ce patron sont considérées comme "servies" et sont exclues du reste de la procédure,

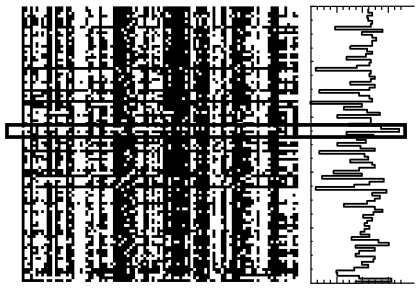
   (e) On recommence l'étape c) pour déterminer le $2^{\text{ème}}$ patron, et ainsi de suite jusqu'à ce que toutes les étoiles aient été servies,

   (f) Si n, le nombre de patrons nécessaire pour satisfaire τ est trop grand, on relâche la contrainte τ. Ainsi chaque étoile admet un plus grand nombre de patrons, et on recommence à partir de b) jusqu'à obtenir $n \leqslant 250$.

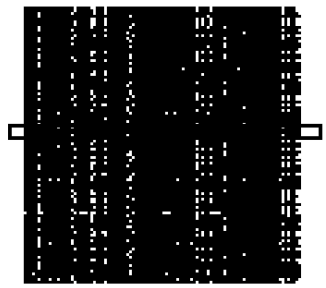
3. Sortie : Le nombre et l'identité des 250 patrons sont connus.



1) La matrice d'acceptabilité *est obtenue par seuillage à la valeur τ de la matrice précédente. Les éléments de cette matrice sont binaires, 1 (en blanc) signifie "cette étoile accepte le masque". Les étoiles sont des colonnes et les masques sont des lignes.*



2) Tri : Le masque totalisant le plus d'étoiles satisfaites est choisi comme premier patron. La somme des lignes est représentée sur la partie droite.



3) Les étoiles servies sont exclues et on itère le 2) pour extraire le second patron.

FIG. 6.12 – *Tri efficace de masques aléatoires.*

Cette méthode est sans biais. En effet, supposons que 12 000 patrons soient autorisés. Si $\tau < 100\%$, la réduction aura lieu et on aura donc moins de 12 000 patrons. Donc la seule valeur possible pour τ est 100% et les 12 000 étoiles n'auront plus qu'un choix : leur propre masque optimisé.

6.7.5 Résultats obtenus

Les résultats évalués sur notre échantillon de 1000 étoiles sont prometteurs. En réduisant d'un facteur 40, la perte se limite à $\leqslant 2\%$. La performance est favorisée par deux effets :

– La perte obtenue est forcément moindre que la consigne fixée. En imposant une consigne $\tau \geqslant 95\%$ on obtient un S/B $\sim 97\%$ car le masque retenu pour toute étoile i satisfait $\tau_i \geqslant \tau$,

– Une conséquence inattendue de l'utilisation de masques aléatoires est qu'elle revient à étendre la recherche semi-exhaustive utilisée pour déterminer les masques optimaux (Llebaria et al. 2002) au prix d'un temps de calcul supplémentaire. Ces essais supplémentaires trouvent parfois de meilleures solutions.

Le degré d'utilisation des patrons est très inégal ; le plus utilisé satisfait à lui seul 130 étoiles (Fig. 6.13). Ce patron le plus "populaire" est de petite surface et sans spécificité, il ne satisfait que les étoiles faibles dont les signes distinctifs sont noyés dans le bruit. On ne doit pas sa présence à un hasard particulièrement favorable : il y en a beaucoup comme lui en compétition sur les mêmes étoiles, dont un seul est retenu. A l'opposé, certaines étoiles brillantes n'acceptent qu'un seul patron.

FIG. 6.13 – *Masque convenant à 130 étoiles. Sa surface est réduite, sa forme standard.*

Bien qu'étant l'hypothèse centrale de notre procédé de réduction, notre critère d'acceptabilité demeure très souple. Nous utiliserons cette propriété à la section suivante pour augmenter la robustesse pratique.

Il existe plusieurs méthodes de tri. Celle qui est proposée est simple et efficace. La solution exacte est hors de portée du calcul et ne serait pas forcément meilleure dans la pratique. L'élément important est la représentativité des étoiles utilisées.

Nous avons envisagé d'autres formes d'acceptabilité, mais le seuillage binaire reste un impératif afin de satisfaire l'objectif scientifique de conserver toutes leurs chances aux étoiles. Le choix d'une préservation moyenne et non minimale du S/B , c'est-à-dire la moyenne des lignes de perte relative (sans seuillage), serait un mauvais critère. L'accumulation d'un grand nombre d'étoiles faibles peut priver une étoile plus brillante de ses masques préférentiels, ce que l'on s'interdit. Seul un seuillage permet d'imposer qu'aucune étoile ne soit délaissée.

6.7.6 Mise en oeuvre de la procédure

Dans la pratique le même jeu de patrons sera appliqué sur plusieurs champs pour réduire les interruptions de télémesure et les risques techniques lors des opérations de rechargement. Nous avons vu au début que la procédure de réduction travaillait à partir d'un regroupement d'étoiles, nommé "champ composite", choisies représentatives des champs stellaires et conditions de mesure qui seront celles de Corot . La population des champs composites se limite à 3 000 étoiles, suivant un compromis avec la puissance de calcul nécessaire. Les caractéristique de ces champs sont les suivantes :

- On choisit une distribution de magnitudes inversée par rapport à la réalité, pour sur-représenter les étoiles brillantes car leurs tâches-image sont plus variées. On utilise quatre, voire huit classes de magnitudes.
- la moitié des étoiles sont issues du centre galactique, l'autre moitié de l'anticentre,
- les imagettes sont simulées avec différents taux de jitter/respiration pour faire face à toutes les valeurs éventuelles.

L'autre facteur important est l'adaptation du critère de qualité globale. L'efficacité de la méthode nous permet de disposer d'une marge de manoeuvre que l'on exploite pour privilégier les rares étoiles brillantes. Nous avons considéré comme important le fait de pouvoir y observer de faibles variations de luminosité, même en l'absence de transits. Un tel choix n'est pas pénalisant pour les étoiles faibles, grâce à la marge de manoeuvre. Pour le mettre en pratique, il suffit de remplacer le seuil fixe τ (cf. Eq. 6.5) par un seuil paramétrable $\tau(m_V)$. On choisit $\tau(16) = 5\%$, qui diminue graduellement jusqu'à $\tau(12) = 1\%$. Malgré cette condition plus stricte, la méthode tient toujours la contrainte des 250 patrons.

Pour accroître le plus possible la robustesse, on contraint encore davantage la méthode en imposant que tout patron de la collection finale convienne au minimum à plusieurs étoiles, dont le nombre dépend de la magnitude (voir Tab. 6.1). Mais cette contrainte supplémentaire est en fait incompatible avec notre démarche initiale où n, le nombre de patrons résulte de la consigne τ. Pour comprendre pourquoi nous considérons le cas extrême (irréaliste) d'une étoile qui n'admet qu'un seul masque, si spécialisé qu'il ne convient à aucune autre étoile. Ce masque doit faire partie de la collection, mais en même temps il ne satisfait pas la nouvelle contrainte. Pour résoudre ce cas on renonce simplement à satisfaire cette étoile, jugée trop atypique pour participer aux patrons génériques.

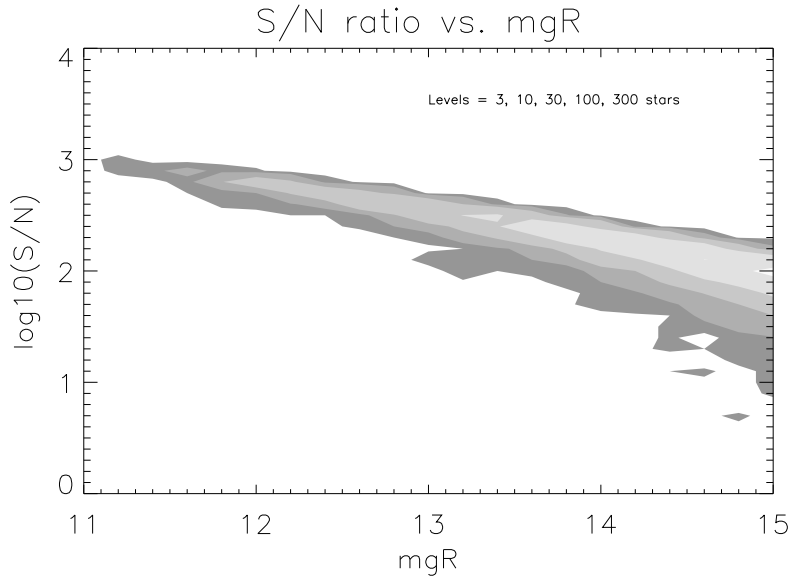


FIG. 6.14 – *Population stellaire en fonction du S/B potentiel. Le niveau de gris indique le nombre d'étoiles d'une magnitude donnée qui ont un S/B donné. Cette figure est établie à partir d'un champ réel. La limite supérieure de la surface indique le cas idéal du bruit photonique pur. Les meilleures étoiles sont les plus rares.*

TAB. 6.1 – *Nombre d'étoiles que doit satisfaire un patron en fonction de la magnitude. Ce nombre imposé est plus important pour les étoiles faibles car elles sont plus tolérantes. On fixe également une nombre maximal de masques par catégorie (n_{masque}) pour réserver plus de masques aux étoiles brillantes. Les étoiles faibles non servies quand n_{masque} est atteint restent exclues de l'échantillon des patrons. Le patron qui leur sera affecté leur fera donc perdre plus de $\tau \times S/B_{opt}$.*

m_V	11.5	12.5	13.5	14.5	15.5
$n_\star$/masque (min)	1	5	15	20	25
n_{masque} (max)	64	64	64	42	16

6.7.7 Conclusion

La précision photométrique de Corot n'est pas compromise par le processus de réduction que nous avons mis au point. La robustesse semble acquise, le S/B obtenu en affectant un jeu de patrons à des étoiles n'ayant pas participé à l'obtention de ce jeu reste satisfaisant, sauf de rares exceptions. On doit la performance de la méthode à l'exploitation de la bonne tolérance de certaines étoiles vis-à-vis de masques désadaptés.

Le jeu de patrons obtenu est une solution globale au problème : il serait néfaste de chercher à retoucher un patron particulier pour l'adapter plus complètement à une étoile donnée. En effet, un patron est un compromis délicat mettant en jeu $\sim$ 40 étoiles ainsi que d'autres patrons. Cette tentative désadapterait d'autres étoiles en nombre bien plus grand.

Un exemple de jeu de patrons est représenté Fig. 6.15. Ils sont groupés par classe de magnitude décroissante à partir de $m_V = 16$. L'anomalie du $6^{\text{ème}}$ patron de magnitude 13 provient sans doute d'un masque élaboré pour une contamination très particulière d'une étoile faible, mais qui ne gêne pas les autres types de contaminations.

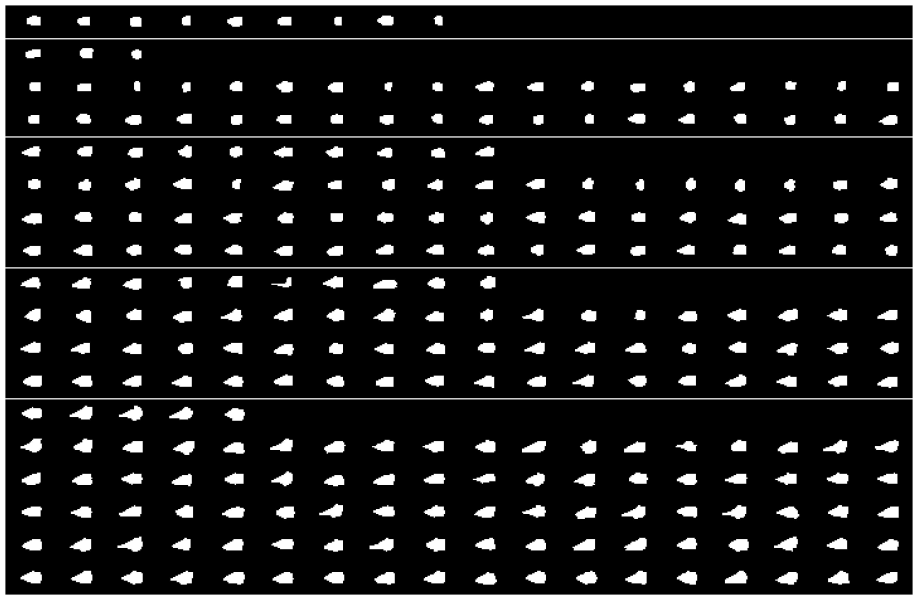


FIG. 6.15 – *Patrons photométriques obtenus avec le processus de réduction développé dans ce travail ; les patrons sont regroupés par classe de magnitude. Les étoiles $m_V = 12$ (en bas) sont plus exigeantes que les étoiles $m_V = 16$ (en haut) très permissives.*

Le nombre d'utilisations des patrons sur un champ (voir Fig. 6.16) montre une décroissance rapide. Les patrons les plus utilisés sont ceux des étoiles faibles. On remarque qu'une vingtaine de patrons restent inutilisés, ce sont ceux qui sont élaborés avec des paramètres trop différents du champ de cibles choisi.

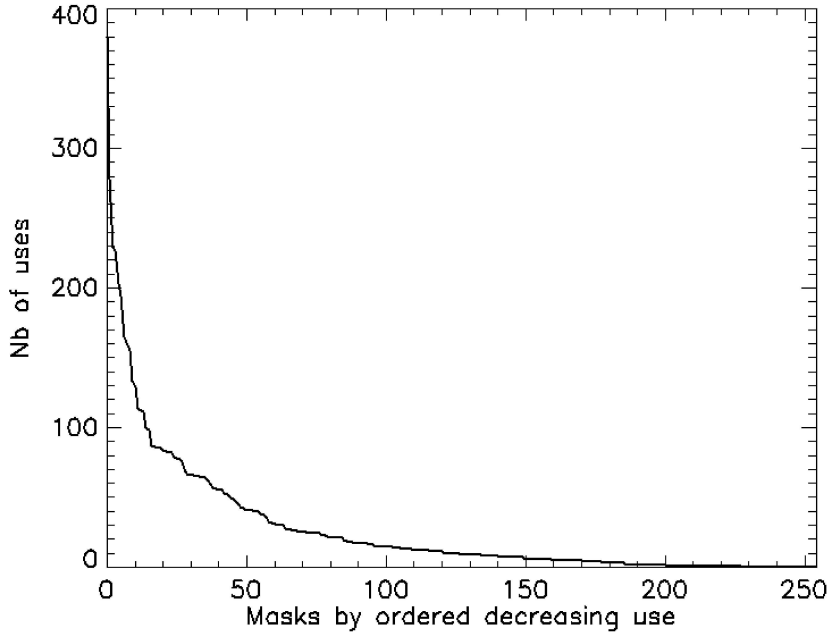


FIG. 6.16 – *Taux d'utilisation des patrons. Le nombre d'utilisation en ordonnée est obtenu sur un vrai champ, après la procédure d'attribution décrite au chapitre suivant.*

6.8 Deuxième publication Llebaria et al. (2003), SPIE.

Photometric masking methods and predicted performances for the CoRoT exoplanetary mission

Llebaria A.[a], Guterman P.[b], Ollivier M.[c]

[b]Laboratoire Astrophysique de Marseille(CNRS), 13776 BP8 Marseille Cedex12, France

[a]Gemplus, BP100, 13881 Gemenos Cedex, France

[c]Institut d'Astrophysique Spatiale(CNRS), Campus Univ. d'Orsay bat.121, 91405 Orsay Cedex, France

ABSTRACT

CoRoT mission for year 2006 is a small space telescope that will measure continuously for 6 months the light flux of 12 000 star in a mission of $2\frac{1}{2}$ years. The aim is to detect small droops in the light curves revealing planets transitting in front of their star. For this, 12 000 logical Regions Of Interest (ROI) are defined on the CCD to optimise each star Signal to Noise Ratio (s/n). Unfortunatly only less than 256 different shapes are permitted for all ROIs, forseeing a loss in global S/N. We found a method wich reduces the 12 000 ROIs to a small set of 250 shapes in a lossless way. Overall performances are discussed.

Keywords: image processing, exoplanets, pattern analysis, stellar photometry

1. INTRODUCTION

CoRoT, standing for Convection, Rotation and planetary Transits, is a European mission led by the CNES* who is the prime contractor.[1,2] Is a main purpose of the orbital telescope CoRoT to detect earth sized exoplanets orbiting at a medium distance around their stars during its 2.5 years mission. The basic method is to find "transits" i.e to detect droops of particular shape in the temporal light curve, corresponding to the tiny brightness decrease when the planet passes in front of its star. Due to planet/star proportions, such droops will be very faint, the detection of 3.10^{-4} in relative brightness decreasing is the expected performance. CoRoT will be in an inertial polar orbit at 850 km permitting to point in the same galactic direction during 6 month runs. During runs the sun is always in CoRoT back side. The indirect light scatered by the earth is reduced by optic means. During runs CoRoT measures continuously light fluxes coming from 12 000 stars selected in its field of view of 3.5 degrees sq. Observing 5 of such areas for 6 months each we estimate that about one hundred planetary systems can detected, along with a dozen of "small planets".[3]

CoRoT is an off-axis 30 cm telescope, with 1200 mm focus and an entrance pupil of 600 cm^2. The optical path is made of 2 afocal mirrors and a 6 lenses dioptric objective. CoRoT works in the $370-950$ nm wavelengths range. On the purpose to discriminate a monochromatic transit from other stellar activity a biprism, just placed before the focal plane, induces a weak dispersion to get somme colored information. The little spectra depending on the star spectral type and class as well as on the position on the field, will be designed in this paper as PFS's. An entry baffle reduces straylight to a 10^{-13} factor. On the focal plane there are 4 CCDs (Marconi MPP) of 2048×2048 pixels of 13.5μm of square size working in frame transfert. To avoid saturations due to bright stars, both CCDs are read every 32 s. Readings are packed by 16 to give a 512 s exposure.

In order to fulfill the telemetry constraints, stars are measured on board and the results are transmited to ground (each exposure is 32 sec long, the sequences of 16 exposures are added on board to form individual measures). These measures constitute a time series or "light curve" for each star. The relative photometry consists on integrating the flux in each ROI (region of interest) which includes the Point Spread Function (PSF) of a selected star. In the instrument working group of CoRoT ROIs are called "masks" and its 2D shapes "patterns". These ROIs are designed from a complex process[4] which aims at maximizing the S/N ratio in view

Further author information: (Send correspondence to Antoine Llebaria)

E-mail: Antoine.Llebaria@obmp.fr, Telephone: +33 (0)4 91 05 59 00

*the Centre National d'Etudes Spatiales (CNES) is the french space agency

of constraints like local crowding, background, jitter, etc. Moreover on board software limits to less than 256 the number of disponible 2D shapes beside other minor constraints. Therefore this shapes will be optimized to be used for large groups of stars depending on place on the field, magnitude, type, etc. In order to define ROIs the full process will be unfold in two steps: in the first one (the "definition" step), a specific ROI for each PSF of selected stars is deduced, in the second one (the "reduction" step) these large sample of ROIs is resumed in a limited set of optimal shapes in compliance with the on-board software requirements.

The definition of the initial (or specific) ROIs has been detailed in a previous paper,[4] instead in the present work we will deal in more detail with the reduction step and the final results. We will discuss in a first place about the optimization criteria, second about the reduction process and last about the predicted performances.

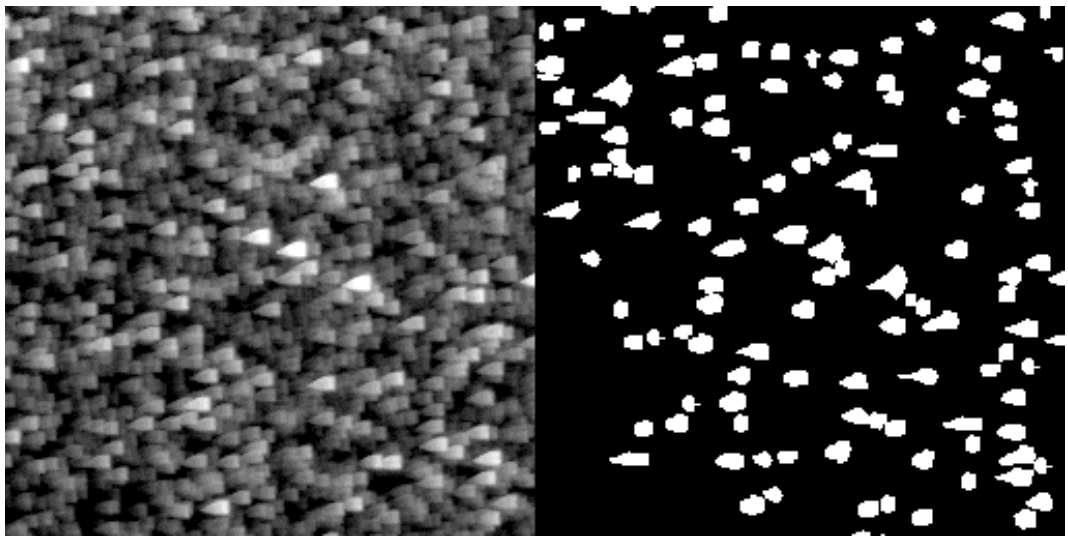


Figure 1. Left: Image 256×256 subfield in logarithmic scale of brightness. Right: correlative ROIs map image

2. OPTIMIZATION CRITERIA

2.1. Defining the noise level

Relative photometry is adequate to detect transit events, therefore the stability of measurements is the crucial point for the exoplanets CoRoT mission. That means to reach the lowest noise level for each serie of measures, therefore the stability of experimental conditions and the low straylight level (mainly due to the earth) are very important items. Because PSFs are different from a star to next one, as we said before, and also because the PSF is subsampled, the only practical photometric method is aperture photometry. The star fields to observe are very crowded and the aperture shape or ROIs will depend on: 1) the PSF distribution, 2) the local background and 3) on the neighboring stars (as matter of fact they belong also to the local background). ROIs are small (~ 70 pixels area) and therefore any tiny displacement or image change in the CCD plane can perceptibily modify the total flux. Displacements are due to residual jitter of the 3D pointing system. Image changes depend mainly on the thermoelastic response of the telescope to the orbital constraints.

The best ROI for each star will minimize such effects to approach the ideal S/N ratio of photon limited noise. Out of operational constraints we will define the optimal ROI applying a exaustive trial and error procedure to determine the shape with the lowest noise to signal ratio for each star. That's why a method to found a simple expression to forecast the noise level for each star and each proposed ROI is crucial.

The signal is defined by the total flux of the star included on the ROI. The noise depends on:

1. photon noise from the star itself;
2. photon noise from the background and the overlapping stars;
3. variable stars neighbours (expected only for $< 0.1\%$ of measures);
4. residual jitter;
5. breathing;
6. periodic (orbital) straylight variations;
7. missed data.

These sources of noise can be grouped on two main classes: a) (Pseudo)white sources and b) quasi-perodic or colored sources. Sources from 1) to 4) correspond to (pseudo)white noise; sources 5) and 6) correspond to colored noise. Even if we can assume a plausible guess for each component we cannot easely mix both classes in a short expression without further assumptions on the future signal processing.

Straylight will show a orbital frequency component (with harmonics and sub-harmonics). Breathing looks much more complex due to the thermic control in flight. In any case such colored noises can be detected and measured because they are global and show a very low frequency cut off. Correcting such terms in the time series of flux data will cancel an important part of the colored noise but a "whitened" residue is certainly foreseeable. This residue will be included in the global expression of noise. Summarizing the discussion of this subjet held in the previous paper,[4] final expression of variance will include three terms: 1) photon noise, 2) jitter noise and 3) whitened noise due to uncorrected residuals of breathing. Obviously variable stars on background, secondary effects of missing data and saturated stars are not considered here. The final expression is:

$$V_{TOT} \quad = \quad F_{TOT} + \delta_\lambda^2 \overline{\Delta g(X)^2} + c_{ES}^2 \overline{g(X)^2} \tag{1}$$

Where V_{TOT} stands for the final variance, F_{TOT} the final flux, X represents the ROI border and $g(X)$ is the flux per unit area in this border. $\overline{\Delta g(X)^2}$ is the quadratic mean of all flux differences in X induced by a series of small PSF displacements, it is the *jitter* term. $\overline{g(X)^2}$ stands for the quadratic mean due to breathing (whitened). δ_λ^2 and c_{ES}^2 are respectively the mean square deviation of jitter and the residual mean square error for breathing.

2.2. Optimization process

This expression has been used to optimize specific ROIs for each star. For each of them an initial ROI is defined assuming a null jitter and breathing noises. In the second step roundROIs are modified to get a minimal variance (relative to the star flux). The third step will reduce the full set of ROIs issued from all selected stars of all selected fields to the limited final set of 256 shapes. In the last step to each star is assigned the most fitted shape of this limited set.

From a practical point of view, we must simulate the 12 000 PSFs (6 000 for each CCD), as well as the full field (including more than 5.10^5 stars) to determine their ROI. PSF are modeled using optical software for 9 positions regularly apart on each CCD, their sampling is 10×10 more fine than the final image, and resumed in a set of 200 monochromatic PSFs in the 350 to 1050 nm wavelenght range (nonuniform) per CCD position. A large set of polychromatic stars is derivated as function of spectral types and classes. Combining this PSFs with a catalog of stars for each field results in a set of 6 000 star images per CCD and a global simulation of both full images of size $2\,048 \times 2\,048$. Once the initial ROI is defined the process deduces a optimal ROI for each of 12 000 stars maximizing the S/N by a quasi exaustive trial and error process. The process deduces S and N from the expected star flux and from the expected local background including other local stars (considered as contaminants). This local background is extracted from the full field simulation of the CCD image.

2.3. Verifiying noise expression with simulations

We use simulations to verify the pertinence of the variance formula we simulate the received signal by moving a PSF across a ROI. The simulation program creates an $\{x_i, y_i\}\, i = 1, \ldots, N$ jitter time serie according to CoRoT parameters. Each second the PSF is shifted by $\{x_i, y_i\}$ and the corresponding star and background fluxes are separatly integrated in the ROI. Table. 1 compares the simulated results with S/N analytic expressions. The table shows a pessimistic estimation using the theoretical expression of noise (by a factor between 1 and 2) relative to the simulations but still, the global agreement is good. Further analysis for the discrepancies is in progress.

Table 1. Computations are made for a 0.86 pixels rms jitter, for 3 stars from strong (Mv=11.3) to mean (Mv=13.9) magnitude. Results are expressed in N/S(ppm) rather than S/N. *SIM* stands for simulation 1D,2D(1) and 2D(2) stands for one dimensional expression, two dimensional expression to first order and two dimensional expression to second order (see Appendix A)

Magnitude Mv	Contamination % rate	SIM ppm	1D ppm	2D(1) ppm	2D(2) ppm
11.3	2%	1463	2404	2646	2625
13.2	4%	784	789	840	830
13.9	6%	1583	2924	2779	2766

3. REDUCTION

To reduce the very large collection of masks ($\sim 100\,000$) to the small set of 2×128 with a minimal degradation in S/N ratio we have tried a set of methods: 1) parametric families, 2) principal component analysis, 3) direct morphing and 4) table sorting. With the last we get, by far, the best results. We design thereafter the group of first ones as *tentative methods* the last one being the *selected method*.

3.1. Tentative methods

3.1.1. Parametric families

Stars, and its respective ROIs, are classed in a multidimensional espace of *a priori* natural parameters: brigthness, position, spectral type and contamination level. They are "traditionnal" major influence factors, *a priori* important. Since ROIs very often change smoothly when the parameters change, the parametric space is divided in a limited set of classes (or families). All stars in a family will be satisfied by a unique shape, obtained by "averaging" all ROIs of this family and thrshoding the result. This method was detailed in a previous paper.[4]

This approach is straightforward but unsatisfactory because: 1) The minimun number of needed classes is larger than the 2×128 allowed shapes, therefore it needs a second reduction step. 2) The pertinent choice of factors is not obvious because underlying parameters, contamination profiles, statistical particularities of the field,... or a concealed combination of some of them can dominate the global S/N. 3) The choice of class limits is somehaw arbitrary, even if it is based on equalization procedures. 4) The compliance of the resulting ROI to convexity requirements is not undoubtely preserved by the process.

3.1.2. Principal components analysis

In this approach we will look for a set of *a posteriori* families. Instead *a priori* (or intuitive) set of factors we will define a minimum set of shapes from the full set of specific ROIs, each one being considered as one multidimensional binary vector. It will help us to define a patterns space where to determine the true dimension of the set of ROIs. Also the S/N cross correlation matrix (see Fig. 2) between stars and ROIs points out to this approach.

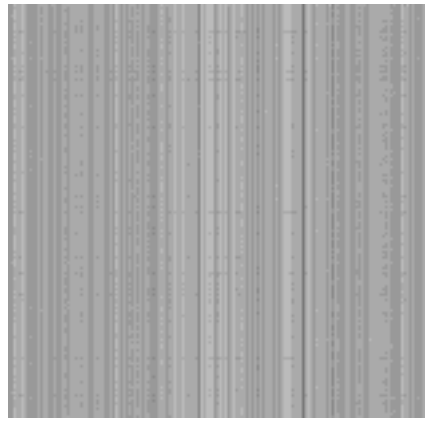


Figure 2. *S/N cross-correlation matrix* between stars and ROIs (presently 1000 stars and its ROIs). Each pixel (i, j) represents the relative S/N for the star i observed with the mask j. Matrix is normalized column by column to the best score (see Fig. 3). Vertical strays mean that bright stars can keep a good S/N with several ROIs. showing that is possible to replace ROIs without further degradation on S/N in the *reduction* step. Faint horizontal dark strays show that a few ROIs are bad for large number of stars.

In spite of some preliminary encouraging results we discard lately this method due to the strong bias : if 1000 shapes where allowed, they would never converge to the 1000 initial ROIs. As in the *parametric families* approach some *a priori* is still present in families boundaries choice. The details are discussed in the Appendix B.

3.1.3. Direct morphing

The *direct morphing* from ROIs to final shapes attempts to suppress previous drawbacks. Using the affine description of ROIs, we run a K-mean clustering algorithm that finds the few shapes that best represent (i.e at a close distance) all ROIs. Due to its mild results this method was discarded. The deep motive is that morphologic similitude is not an adapted metric for S/N optimisation. Details are discussed in the Appendix C.

3.2. Selected method

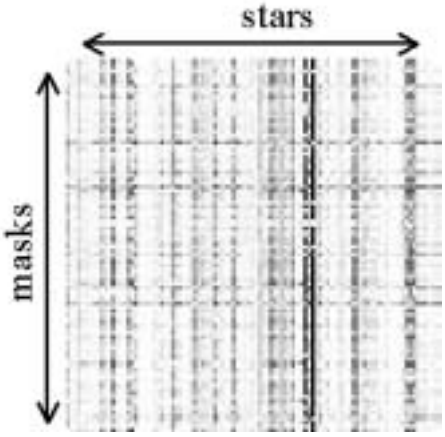
3.2.1. Table sorting basics

The selected method, i.e. *table sorting* is directly S/N oriented, that's why its results are clearly better than previous ones. As matter of fact the previous methods build generic shapes in spaces defined by diverse features close linked to S/N (brigthness, ROIs etc..), not by the loss of S/N itself. The matrix of cross S/N ratios (see Fig. 2) shows that for a large set of ROIs the S/N is very often close to the best S/N i.e. the S/N of individual stars is tolerant relative to a panel of less of adapted ROI. Moreover because the search for an optimal ROI can not be exhaustive in the S/N matrix "foreings" ROIs can show better performance. Starting from this remark, our strategy is to find few ROIs that are *acceptable* for as much stars as possible. To deal with the trade off between a good global S/N and a reduced number of shapes we define a convenient acceptability criterium for a couple {star, ROI} : the maximum acceptable S/N loss ratio τ. Indeed 1) τ is relative to each star 2) if all stars satisfy τ then the global S/N also satisfies τ.

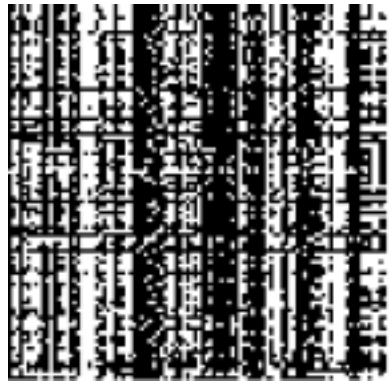
3.2.2. Strategy

Practically we consider all ROIs found in the set of specific ROIs as a tank of random ROIs. Given a τ, our process determines the required number of shapes to satisfy τ. By successive tries, we adjust τ to obtain the 2×128 shapes. 2 steps are necessary :

1. Convert the exaustive matrix of fig. 2 into the S/N losses matrix of fig. 3
2. The best ROI, i.e. the row on the acceptability matrix with a maximal number of compliant couples, is removed from the matrix as well as the compliant stars. A new losses matrix with the remaining stars and ROIs is defined. Iteratively convert to shape "status" the ROIs acceptable for the largest number of compliant stars, as described Fig. 3

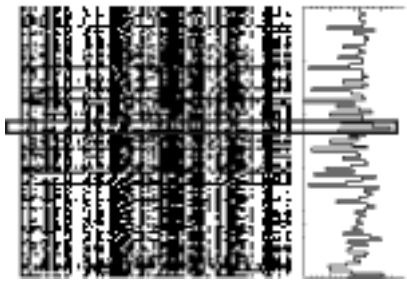


The *losses matrix* results from the S/N matrix in fig. 2. Each column of the S/N matrix has been normalized to its highest column value. Columns show the S/N loss ratio across ROIs.

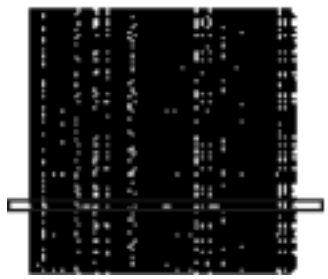


The *acceptability matrix* i.e. the tresholded *losses matrix* to loss τ. White spots signal compliant couples of $\{\text{star}_i, \text{ROI}_j\}$ to this loss level

Figure 3. The *losses matrix* and its derivated *acceptability matrix*



In the *acceptability matrix* the most white spotted line marks the best shape to extract. This shape is assigned to all compliant stars of this line



Compliant stars for the last extracted shape are switched off. A new iteration can start in order to find a next *best* shape and so until star exaustion

Figure 4. Iterative procedure to determine the set of shapes from the S/N *acceptability matrix*

3. The 1) and 2) steps are iterated until star exhaustion, or a maximum of 128 ROIs (per CCD). If the maximum number of ROIs is reached without exhausting all stars, we have to decide the increase of τ losses for a subset of stars or admit a more reduced S/N for the remaining stars. Trials with actual fields shows a very small proportion of remaining stars for $(1-\tau)$ S/N loss levels lower than 10% starting from more than 120 000 stars-ROIs pairs (from different fields but relative to the same CCD).

At the end we obtain:

1. The shapes list.
2. The guarantee that local S/N losses are better than $(1-\tau)$.
3. Temporarly a shape (or "generic mask") is assigned to each star referred in the S/N matrix.

Because in the actual process only a subset of stars is involved in the S/N acceptability iteration, in the final run for each star the shape showing the best S/N is assigned as its definitive ROI.

With a test deck we reduced 1000 stars and specific ROIs to 26 shapes. The S/N loss $(1-\tau)$ fall under 1%. Hence the global S/N is close to the (dedicated ROI)/(best random ROI) ratio. This one is lower than 100% because the preliminary ROI optimizations are only semi-exhaustive. Therefore we can consider our reduction method as lossless. However practical issues dealing with many thousands of stars show somehow less favorable rates (see section 4).

3.2.3. From theory to practice

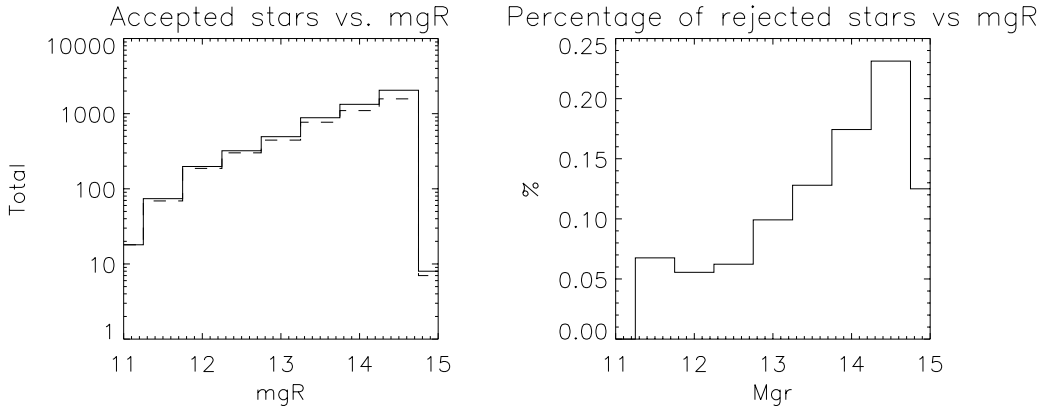


Figure 5. *Left* Accepted stars per magnitude in a typical field. *Right* Size of accepted stars overimposed to the acceptance histogram. Small masks are de most frequent ones.

Generic masks will be optimized for a large set of observable fields, not just for only one of them. Each field contains as much as $2 \times 6\,000$ potential stars to observe, therefore the exaustive analysis with all candidates ($> 50 \times 2 \times 6\,000$ candidates) requiring an exaustive S/N acceptability matrix of $600\,000^2$ elements is clearly out of our reach. Instead a stochastic selection has been applied. This process:

1. selects at random a large sample of stars (> 3000) between all candidates in a large sample of fields and builds the S/N acceptability matrix,
2. finds the limited set ($< 2 \times 128$) of the most common accepted shapes (or "generic masks") following the *table sorting* method.

A final procedure assigns ROIs in each field to 6000 star candidates from the limited set of shapes. Once shapes assigned, a small proportion of candidates ($< 10\%$) are rejected due to practical constraints like to avoid: 1) the intersection with borders and 2) the overlapping between ROIs in the CCD. (see fig. 5).

4. PERFORMANCE

Performance has been mesured in a set of 12 fields in the Milky Way of $1.6° \times 1.6°$ each. The number of stars per field is more than 120 000 to mgR 19.6. They are more than 6 000 candidates in each field to mgR 15. Sets of 2×128 shapes have been extracted from this fields, and ROIs and its positions defined for most of them. We present here the typical results for one of these fields:

- The 2D histogram of the star number as function of S/N and magnitude.(fig. 6)
- The histogram of total ROIs per shape or frequency use in decreasing order of use from left to right. (fig. 7 left)

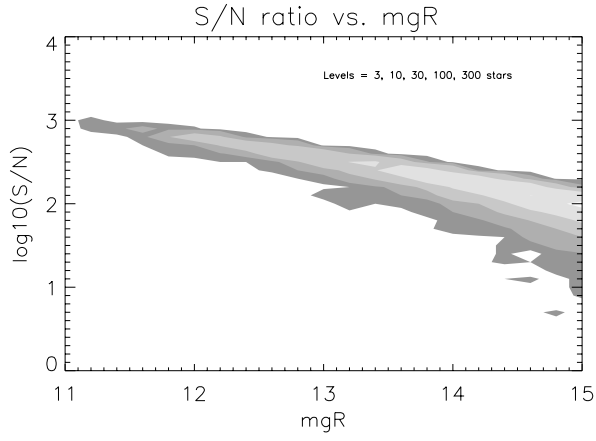


Figure 6. Number of stars per magnitude and S/N ratio (scales in $\log_{10}(S/N)$ vs. mgR coordinates)

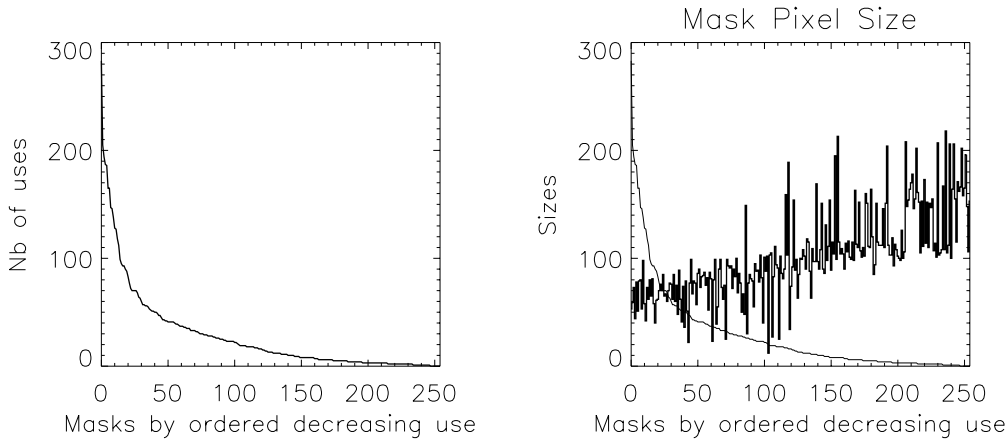


Figure 7. *Left* Accepted stars per magnitude in a typical field. *Right* The graph showing the size of masks is overimposed to the graph of accepted stars per magnitude

- Shape sizes overimposed to the frequency use.
- The histogram of shape sizes (fig. 7 right)
- The ROI Statistics: 1) Size histogram, 2) 2D histogram of size and magnitude, 3) 2D Histogram of length and magnitude and 4) 2D histogram of height and magnitude (fig. 8)

The 2D histogram (fig. 6) of the star number as function of S/N and magnitude (mgR) shows a dense population of stars grouped along the theorical limit of S/N rate as fuction of magnitude (scales in $\log_{10}(S/N)$ vs. mgR coordinates, the Poisson noise limit is a straight line of slope -0.2 per 1 mg). In the graph, the distribution ridge is shifted of 0.2 in $\log_{10}(S/N)$ units, corresponding to the $\sim 20\%$ decrease in S/N performance relative to the pure Poisson noise of the star PSF without background. Such degradation includes the noise due to contamination from neighbour stars and from the background level, the jitter effects, etc.

The ROIs statistics (fig. 8 left) shows: 1) the size distribution of ROI shapes (mean of 70 pixels), 2) the frequency of use: most of stars are observed with ROIs of < 100 pixels, 3) the height and length distributions:

mean of 9×15, 4) 2D histograms of nb of ROIs as function of magnitude and size, low magnitudes use big ROIs as expected.

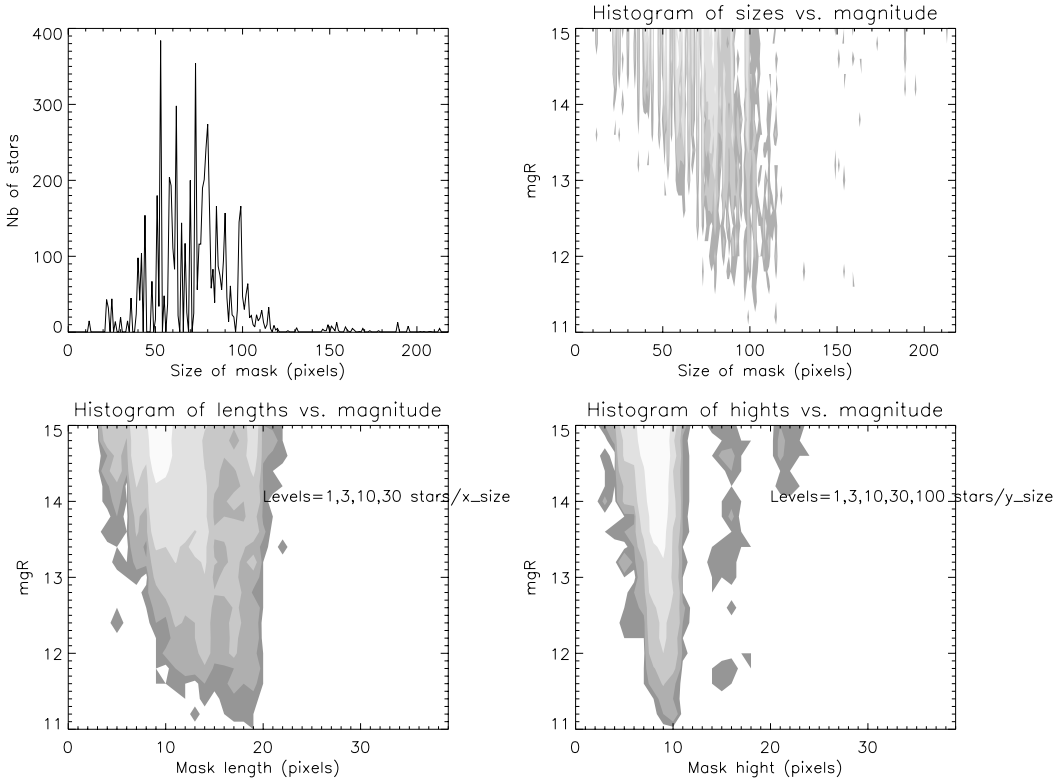


Figure 8. ROI Statistics: 1) Size histogram, 2) 2D histogram of size and magnitude, 3) 2D Histogram of length and magnitude and 4) 2D histogram of height and magnitude

5. CONCLUSION

This paper has presented the procedure to obtain the reduced set of masks used into photometric integration in the CoRoT exoplanets experiment as well as the expected performances. Operational constraints to a limited set of masks. In fact a limited subset of masks (< 50) are enough to measure a majority of stars with a low degradation in the S/N ratio. The procedure described here shows an outstanding improvement relative to the classical methods of reduction of pattern sets. The expected overall performance is better than 80% of the "photon noise" limited observations and the field crowding will discard in mean less than 25% of less bright stars. Future work will try to verify this performances with the full set of CoRoT exoplanets fields.

5.1. Acknowledgments

We are grateful to E.Brier for his helpful discussions about the jitter statistics as well to M.Auvergne for the fruitfull exchanges on mask determination and noise level determination. This work has been found by the CNES (the french space agency) and the CNRS.

APPENDIX A. 2D ANALYTIC S/N EXPRESSIONS

Complete expressions of S/N with 2D PSF and jitter can be approximated by a 1^{st} or a 2^{nd} order formulas. $F(i,j)$, $F_{TOT}(i,j)$ and $\lambda(i,j)$ are 2D expressions for distributions of PSF flux, total flux and (i,j) positions

respectively. i and j stand for continuous pixel coordinates. ROI is the window area. $F_{TOT} = F + F_{Background}$. The $\lambda(i,j)$ distribution of (i,j) can be defined by the set of its centered moments $\mu_{20}, \mu_{11}, \mu_{02}, \ldots$. Let be $< F >= \iint_{ROI} F * \lambda$ and $< F_{TOT} >= \iint_{ROI} F_{TOT} * \lambda$ the mean fluxes for the star only and for the full window respectively. Let be $q_1 = \oint \overrightarrow{F_{TOT}}.\overrightarrow{i}$ (resp. $q_2 = \oint \overrightarrow{F_{TOT}}.\overrightarrow{j}$) that is to say differences between vertical (resp. horizontal) borders.

S/N expression at 1^{st} order is :

$$S/N = \frac{< F >}{\sqrt{< F_{TOT} > + \mu_{20}^2 q_1^2 + \mu_{02}^2 q_2^2 + 2\mu_{11} q_1 q_2}}$$

Let $q_3 = \oint \overrightarrow{\text{grad}} F_{TOT}.\overrightarrow{i}$ (resp. $q_4 = \oint \overrightarrow{\text{grad}} F_{TOT}.\overrightarrow{j}$) thats to say differences between horizontal (resp. vertical) slopes. Noise expression of noise contribution becomes at 2^{nd} order :

$$\begin{aligned} N^2 &= < F_{TOT} > + \mu_{20}(q_1^2 + q_3/2) + \mu_{02}(q_2^2 + q_4/2) + 2\mu_{11} q_1 q_2 \\ &+ \mu_{02}^2 q_3^2 + \mu_{02}^2 q_4^2 + \mu_{20}\mu_{02} q_3 q_4 \\ &+ \mu_{30}\, q_1 q_3 + \mu_{12}\, q_1 q_4 + \mu_{21}\, q_2 q_3 + \mu_{30}\, q_2 q_4 + \mu_{40}\, q_3^2/4 + \mu_{22}\, q_3 q_4/2 + \mu_{40}\, q_4^2/4 \end{aligned}$$

APPENDIX B. PRINCIPAL COMPONENTS ANALYSIS

Instead of classification based on parameters what we "feel" important, we work here only with the visible effect denoted on ROIs. We apply the *a priori* method, properly formalised, on the PCA analysis. To isolate ROIs caracteristics factors we find the eigenvalues of the covariance matrix of the ROIs vector coordinate. The first 15% of coordinates account for more of 75% of ROI variations. We keep them as *a posteriori* major factors.

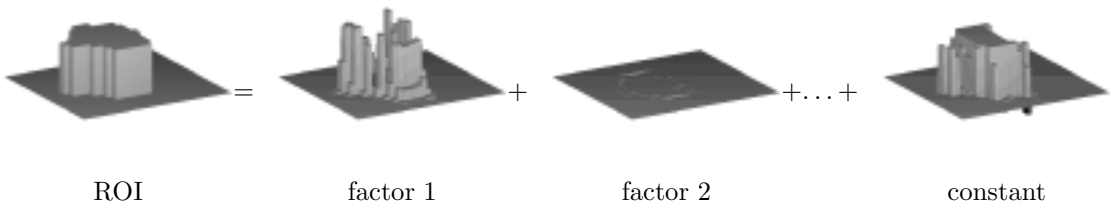


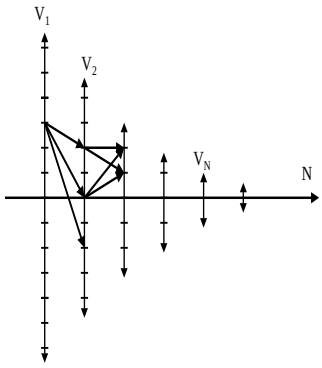
Figure 9. Example of ROI decomposition in independant factors.

In this process the columns of an ROI image (37×16 pixels) are stacked in a vector $\overrightarrow{m}$ with one binary coordinate per pixel in the canonic base (i.e 592-coordinates). A coordinate of $\overrightarrow{m}$ set to 1 signifies the corresponding pixel is used in the ROI. For shake of clarity let us fix the number of ROIs to 1 000. The $\{\overrightarrow{m_i}, i = 1 \cdots 1000\}$ induce a sub-space where we assume the searched shapes to reside. We searching the underlying dimension of the system, we cannot find more than 100 independent $\{\overrightarrow{m_i}\}$. This means that 492 of the 592 pixels actually never toggle or are tied with another. So we change for this new base of 100 coordinates. Note that such coordinates are no more binaries.

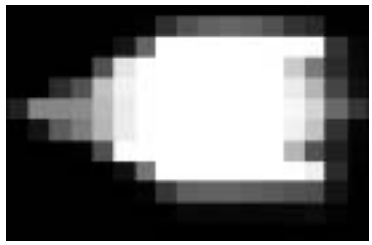
Influence factor formalization : A combinaison of coordinates whose variations are *not correlated* to the others. To determine them we consider the covariance matrix A of $\overrightarrow{m_i}$ coordinates. Non-diagonal terms are covariance between coordinates (i.e factors). Let's diagonalize A and change to the base of eigenvectors $\{\overrightarrow{V_i}, i = 1 \cdots 100\}$. The non-diagonal terms of A are null so the new coordinates are independant. In other words, $\{\overrightarrow{V_i}\}$ are our influence factors. We formalize thereafter all this concepts

- *Major factors*: Those who varies most between ROIs. The variance of the i^{th} coordinate is the eigenvalue λ_i. So we reorder $\{\overrightarrow{V_i}\}$ by descending λ_i.

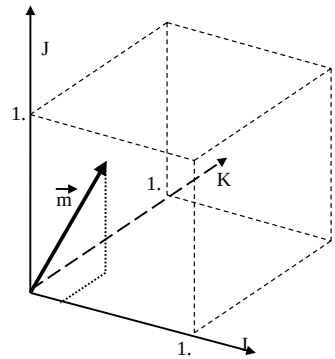
- *Family*: All members of a family share a close value for a factor. Analytically this value is the projection of an $\overrightarrow{ROI}$ on a $\overrightarrow{V_i}$. To share into 250 families, we cut factors values into equal parts or equal populations of ROIs, as shown Fig. 10 left.
- *Rebinarization*: The obtained shapes (see Fig. 10, center) are back translated to the canonic base. As it is, shapes a negligeable chance to show binary coordinates. A way to solve this is to notice that binary shapes are located at corners of an hypercube in the continuous space (see Fig. 10, right). This justifies the intuitive choice made in **Sect. 3.1** of pixel averaging among a family.



Significant factors or eigenvalues are quantified



The resulting best shape is almost always a continuous one



Continuous shapes are included in a cube of binary shapes

Figure 10. From continous shape to the closest binary shape

APPENDIX C. DIRECT MORPHING

We use a progressive morphing from ROIs to shapes. We assume that 2 *similar* shapes applied to the same star will give a *similar* S/N. In this section it is convenient to represent shapes by affine points with binaries coordinates. Logically, 2 shapes $m_i\, m_j$ are said similar if they have only few differing pixels. In other terms, if the distance $\|\overrightarrow{m_i\, m_j}\|$ is small. The term of data clusterring algorithms is, given a numerous set of points m_i, to find a limited number of representative catergories. Categories are resumed to their centers c_j. The algo finds the c_j and assigns m_i accordingly such as $\sum_j \sum_i \|\overrightarrow{c_j\, m_i}\|$ is minimal.

C.1. Basic algorithm

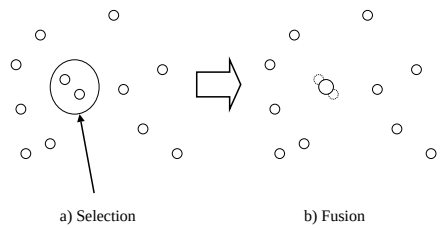


Figure 11. At each iteration, the closest pair of type {point, point} or {point, centroïd} or {centroïd, centroïd} is replaced by its centroïd, hence reducing by one the number of weighted points.

A simple algorithm doing this is illustrated in Fig. 11. The principle is, starting from the complete set of points, we remove one point at each iteration.[5] Elimination rule : the 2 closest points are replaced by their barycenter (centroïds). Advantages : 1) No more *a priori*. 2) Shapes converge obviously throught ROIs as their number increase.

The algorithm iterates $nROI - nShape$ times. The complexity of all distances computation is heavy, roughtly in $\frac{n^2 \times \text{nb iterations}}{2}$, but can be significantly reduced, for instance by sharing the space into virtual boxes and computing distances only inside and between boxes. The algo is completed by the same rebinarisation than App. B. The result of this basic algorithm converges badly. Indeed, all points condense on the smallest ROI. This seems due to that points with few non-null coordinates have the more chances to be close together since both are close to the origin. Consequently central points start to condense and atract all the others.

C.2. K-means problem

Its general term is : Given n pupils spread over a region, where to place K schools, $K < n$ such as the total distances to schools is minimun. In our case, pupils are ROIs and schools are shapes. There is no analytic answer, but a class of iterative algorithms[6] known also as K-means cluster algorithm. The computational complexity falls downto $K \times n$ per iteration. Lets detail one. Initial state : Spread K seeds (future shapes) over the space. We preferently choose for seeds K regulary distributed ROI to respect convergence toward initial ROIs as shapes number increases. Figure. 12 illustrates the 2 steps of an iteration. The iteration stops when a

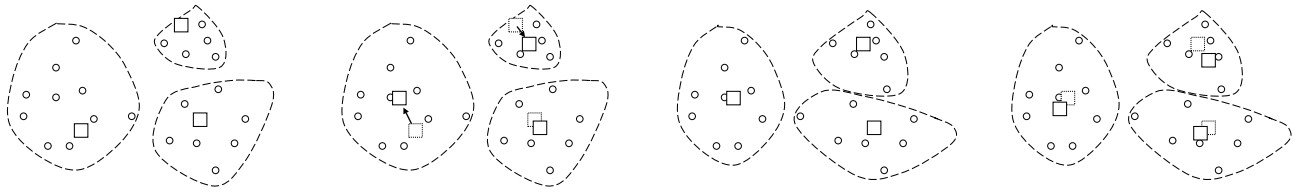


Figure 12. At each iteration, all points are assigned the closest seed, then the seed is moved to the cluster's centroïd

convergence criterium is reached, like no more decrease of the global distance. Centroïds are then transformed to shapes by binarisation (App. B). The result we obtained is satisfying, only 1.7% surface differ between initial ROIs and final shapes. Moreover this ratio is quite independent from initial seed choice and the algorithm converges in a few iterations. Unfortunatly, we note that the basic hypothesis saying that close ROI would give close S/N was non relevant. In fact, studying the relative distance versus S/N degradation, we note that same distance could lead to quite different S/N. We conclude that morphology criterium is risky and we abandon morphing methods.

REFERENCES

1. D. Rouan, A. Baglin, P. Barge, E. Copet, M. Deleuil, A. Léger, J. Schneider, D. Toublanc and A. Vuillemin "Searching for exosolar planets with the CoRoT space mission", *Physics and Chemistry of the Earth Part C*, **24, 5**, pp. 567–571, 2000.
2. D. Rouan, A. Baglin, E. Copet, J. Schneider, P. Barge, M. Deleuil, A. Vuillemin and A. Léger, "The Exosolar Planets Program of the CoRoT satellite", *Earth, Moon, and Planets*, **81, 1**, pp. 79–82, 2000.
3. P. Bord, D. Rouan, A. and A. Léger, "Exoplanet detection capability of the COROT space mission", astro-ph/10305159 *A&A*, in press, 2003.
4. A. Llebaria, A. Vuillemin, P. Guterman, P. Barge, Proc.SPIE **4849** # 112, 2002.
5. Everitt, Brian S., Cluster Analysis, Arnold Publications, 2001.
6. A.K. Jain and R.C. Dubes, Algorithms for Clustering Data. Englewood Cli s, NJ: Prentice Hall, 1988.

Chapitre 7

Attribution des patrons sur une image complète

Nous avons vu à la section précédente comment déterminer un nombre réduit de patrons. Il reste à les répartir de manière optimisée sur les cibles du champ d'étoiles sélectionné. Il ne s'agit pas simplement d'attribuer à chaque étoile son meilleur patron possible. Lors du calcul des patrons, les étoiles étaient prises en compte indépendamment les unes des autres. Mais en fait, les contaminants d'une cible sont souvent eux-mêmes des cibles. N'ayant pas la possibilité de se chevaucher leurs patrons peuvent se trouver en conflit. Ce sont des situations qu'il faut gérer, les "collisions", interdites pour des raisons techniques, pouvant aboutir à un rejet pur et simple de certaines cibles potentielles.

Pour cette optimisation nous nous trouvons à nouveau devant le dilemme de devoir arbitrer entre qualité et quantité de l'information disponible sur les CCDs. Ce besoin d'optimisation est cependant plus simple à gérer que précédemment (cf. §6.7.1). Les deux antagonistes sont ici le S/B individuel qui augmente avec la surface du patron (donc son encombrement) et le nombre d'étoiles rejetées à cause des collisions.

Après avoir testé les deux alternatives opposées sous la forme d'une procédure privilégiant le S/B individuel et d'une autre réduisant la surface des ouvertures, j'ai mis au point une méthode très simple qui résout le dilemme de manière avantageuse. Elle consiste en une attribution itérative suivant la priorité des S/B , combinée à un repêchage "équitable" en cas de collision. Cette solution évite d'avoir à choisir un compromis car elle se révèle supérieure aux deux autres possibilités, chacune sur son terrain de prédilection. C'est-à-dire que le S/B total est supérieur à celui obtenu par la procédure orientée S/B et le nombre de cibles retenues est plus grand qu'avec la procédure qui réduit la surface.

Dans la suite, le terme *fenêtre* désignera un patron positionné sur une étoile. Il y a donc 250 patrons et 12 000 fenêtres. Les contraintes techniques sont les suivantes :

Collisions : deux fenêtres ne doivent avoir aucun pixel en commun ;

Colonne inerte : aucune fenêtre ne doit chevaucher l'inter-colonne centrale située entre les colonnes 1024 et 1025 du CCD (en commençant à 1), à cause de contraintes propres à l'électronique de mesure ;

Bordures : les fenêtres ne doivent pas "mordre" au delà du bord externe du CCD ;

Saturations : on s'interdit de placer une fenêtre sur les traînées de saturation, impropres à la mesure.

L'algorithme mis au point étant destiné à être intégré dans une chaîne opérationnelle facile à gérer, il a également fallu soigner sa réalisation et le rendre simple. Pour cela je l'ai épuré et rationnalisé en cloisonnant les flux de données et en centralisant les données d'interface. L'algorithme est également documenté et produit des comptes rendus d'éxécution.

7.1 Méthodes testées

Les étapes d'attribution sont les mêmes quelle que soit la méthode utilisée. Elles procèdent par champ d'un CCD complet contenant 6 000 cibles. Les cibles désignées ne sont à ce stade que des candidats. Pour en retenir 6 000 par CCD, il faut en disposer d'environ 15% de plus. Voici le schéma commun :

- Les données d'entrée sont la collection de patrons, la PSF de chaque cible, une imagette de sa contamination, l'image du champ complet incluant l'interpénétration des cibles, les saturations et le traînage,
- on calcule une matrice signal à bruit des 250 patrons disponibles appliqués aux $\geqslant$ 6 000 cibles,
- les étoiles sont classées par ordre de S/B ,
- la boucle d'affectation est au coeur de la méthode. Sa description figure plus bas,
- les données de sortie sont les positions et patrons des fenêtres, ainsi que des informations complémentaires destinées à la détection des transits.

Voici les deux boucles d'affectation de spécificité opposées utilisées initialement.

Méthode 1 : elle est séquentielle, orientée S/B . Les étoiles sont affectées par ordre de S/B décroissant. Chacune reçoit son meilleur patron possible. En

cas de collision, on passe à la suivante. L'étoile est donc perdue, au profit de sa rivale déjà en place qui offre un meilleur S/B .

Méthode 2 : elle est séquentielle, orientée S/B avec compromis de surface. Les étoiles sont affectées dans le même ordre que précédemment, mais ne reçoivent pas leur meilleur patron. Elles reçoivent le plus petit des patron qui leur cause $\leqslant 5\%$ de perte (seuil ajustable). Les fenêtres étant plus petites, le CCD en contient davantage.

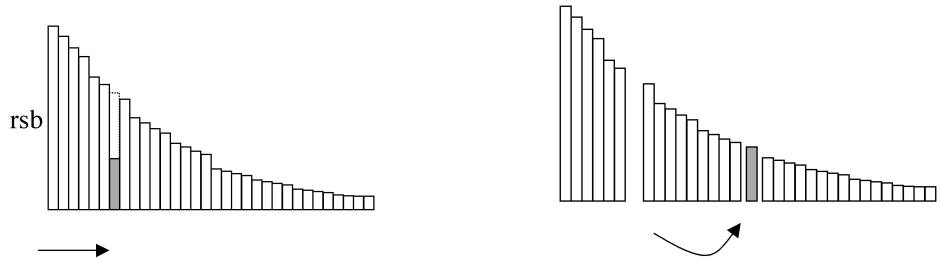
Les défauts de l'une et l'autre de ces techniques sont qu'il reste souvent $\leqslant 6000$ cibles retenues, parfois privées d'étoiles dont le potentiel en S/B était pourtant intéressant, ou bien que le S/B est systématiquement diminué.

7.2 Méthode choisie

Chacune des méthodes précédentes donne un résultat sous forme d'un couple $\{q_1, n_1\}$ (resp. $\{q_2, n_2\}$), où q est le S/B total et n ne nombre de cibles retenues. q et n évoluent en sens contraire : $q_1 > q_2$ et $n_1 < n_2$. On ne peut donc pas dire objectivement quelle méthode est supérieure à l'autre. Dans de tels cas on est en général contraint de relier q et n par classement, pondération, seuil ou toute autre classe d'équivalence pour se ramener à un critère unique. Mais ce choix comporte une part d'arbitraire.

La seule manière de départager en toute objectivité est de trouver une méthode dont le résultat $\{q_3, n_3\}$ satisfasse à la fois $q_3 > q_1$ et $n_3 > n_2$. C'est ce que réalise la méthode mise au point, que l'on qualifie d' "itérative avec repêchage équitable". Son principe est simple : en l'absence de collision une étoile reçoit son meilleur patron, de même que dans la méthode orientée S/B . Mais en cas de collision elle n'est pas éliminées tant que son S/B potentiel reste intéressant. On exploite pour cela la forme du patron grâce à un algorithme itératif qui conserve une combinatoire réduite.

Le fonctionnement est très simple. C'est une gestion de file d'attente (voir Fig. 7.1). Les étoiles sont rangées par ordre de S/B décroissant tous patrons confondus. La première étoile reçoit son meilleur patron (donc le même qu'avec la méthode 1). Puis la seconde etc.. Lors d'une collision, l'étoile est privée de ce patron. Son S/B potentiel est recalculé avec les patrons restants et elle est réinsérée en file d'attente au rang correspondant. Ainsi les étoiles qui conservent un bon S/B potentiel même une fois dégradé, sont réessayées en priorité. C'est en cela que le "repêchage" est dit "équitable". C'est seulement quand tous ses patrons sont épuisés qu'une cible est éliminée ; mais à ce stade elle est déjà fortement dégradée.



Les étoiles sont classées et affectées par ordre S/B décroissant. Si une étoile collisionne, elle est privée de son meilleur patron. La colonne grisée représente le meilleur S/B de cette étoile avec les patrons restants.

L'étoile est replacée dans la file d'attente au rang qui lui correspond à son nouveau "meilleur S/B ". L'affectation se poursuit en reprenant le processus là où il s'était interrompu.

FIG. 7.1 – *Attribution priorité S/B avec repêchage*

Les résultats des trois méthodes sont comparés table 7.1. La méthode itérative conserve le même S/B individuel sur toutes les cibles que la méthode 1 a retenu aussi. Elle permet en plus de distribuer des patrons supplémentaires avec un bon niveau de S/B . Le nombre d'étoiles perdues se limite à 10% et il s'agit des moins intéressantes. Un examen plus précis montre que les cibles perdues peuvent être : (a) des cibles recouvertes par le patron de cibles plus brillantes, (b) des cibles en contact avec une étoile brillante ou avec la bordure du CCD (masques interdits). Ces cas ne posent pas vraiment problème. Les premières sont en général très contaminées, les secondes perdent une part importante de leur flux qui, sommé dans la fenêtre de leur voisine, ne pourra pas en être distingué.

Pour vérifier cette analyse sur tous les cas nous avons ajouté quelques masques qualifiés "de sauvetage", parce qu'ils sont dessinés pour tester les cas spécifiques précédents sans souci du S/B . Ils ont la forme d'une pixel unique ou d'un disque plein évidée d'un secteur de 25%. Les étoiles restantes ont alors toutes trouvé un patron, ce qui confirme que leur précédente élimination était due à la géométrie du champ, et que la procédure ne laisse pas perdre d'étoile intéressante.

7.3 Résultat

Nous avons comparé les méthodes sur une champ-test qui compte 5 193 cibles candidates. Le résultat se trouve table 7.1. La méthode 1 orientée S/B perd 1 000 cibles mais les cibles restantes conservent leur meilleur niveau de S/B . En réduisant la taille des fenêtres, la méthode 2 affecte 200 cibles supplémentaires

au prix de 3% de S/B perdu. La méthode 3 conserve au moins les mêmes cibles que la méthode 1 avec un S/B égal, mais en plus elle regagne 200 cibles sur la méthode 2 et à moindre perte.

TAB. 7.1 – *Comparaison des méthodes. Le S/B de la méthode 1 est la référence utilisée pour faire la comparaison.*

Méthode	1	2	3
Priorité	S/B	n cibles	S/B itérative
fenêtres	4100	4300	4573
ratio S/B	-0%	-3%	-0% (les brillantes)
	(référence)		$\leqslant$ -1% la plupart des autres

La table 7.2 présente le résultat d'une affectation pour un champ encombré du centre galactique. Les 7225 étoiles candidates sont plus nombreuses que les 6 000 cibles possibles. On voit qu'avec une réserve de candidat, la procédure produit 95% des courbes de lumière qui perdront $< 10\%$ de S/B par rapport à leur meilleur patron possible. L'autre avantage est que la perte reste cantonnée aux étoiles qui avaient déjà une carence de S/B .

TAB. 7.2 – *Extrait du compte rendu d'exécution du programme d'affectation joué sur un champ de 7 225 candidates du centre galactique. Il indique le potentiel en étoiles et en S/B , le nombre de cibles affectées avec différents taux de perte, et la surface totale couverte par les fenêtres.*

```
; affecting windows to stars Tue Apr 26 11 :00 :20 2005
; computing sn and reordering affectations Tue Apr 26 11 :00 :22 2005
; potential                    :  7225 stars ,  1.120E+06 total sn
; affected targets             :  5563,         with sn loss <1% of best
;                                 175,          with 90%<sn<=99% of best
;                                 194,          with 50%<sn<=90% of best
; unaffected or lost > 50% sn  :  68,
; masks area ratio             :  11.5%
```

La table 7.3 montre sur un exemple que la stratégie d'affectation parvient à préserver préférentiellement les étoiles de S/B $\geqslant 100$ (sur 8.5 min) qui reçoivent presque toutes leur meilleur patron possible. Elle reste très efficace dans les tranches inférieures où plus de 96% des étoiles de S/B $\geqslant 40$ ne subissent aucun dommage. L' essentiel de l'effet des collisions est absorbé par les étoiles faibles S/B < 25. Au total plus de 90% des 6 000 canaux ont reçu une étoile.

TAB. 7.3 – *Taux de préservation du S/B en fonction S/B pour la plupart des 5 564 étoiles attribuées d'un champ du centre galactique.*

S/B	398.1	251.2	158.5	100.0	63.1	39.8	25.1	15.8
τ								
40.%	0	0	0	0	1	1	3	6
50.%	0	0	0	0	1	2	10	18
63.%	0	0	2	2	3	10	29	38
79.%	0	2	0	7	27	42	74	40
100.%	72	210	411	715	1083	1513	892	171

La table 7.4 montre l'effet cumulé de la réduction et de l'attribution. On voit que les étoiles plus brillantes que $m_V = 14$ ont des S/B relativement regroupés, signe d'une bonne préservation à travers les procédures. Pour $m_V \geqslant 14$, l'étalement des S/B reflète de la priorité inférieure de ces étoiles à travers les différentes étapes. Elles servent en quelque sorte de "régulateur" pour préserver les meilleures candidates.

TAB. 7.4 – *S/B sur 8.5 min en fonction de la magnitude pour la même attribution que la figure.7.3.*

m_V	11.5	12.0	12.5	13.0	13.5	14.0	14.5	15.0	15.5
S/B									
15.8	0	1	1	4	2	6	16	56	188
25.1	0	0	1	1	8	15	49	167	768
39.8	0	0	5	8	11	45	160	614	728
63.1	0	1	4	9	40	171	507	347	38
100.0	0	6	7	40	191	356	113	11	0
158.5	2	6	53	169	150	27	4	1	0
251.2	4	43	101	55	8	1	0	0	0
398.1	22	35	8	1	1	0	0	0	0
631.0	3	1	0	0	0	0	0	0	0

7.4 Double critère de priorité

Il est nécessaire de prendre en compte un deuxième critère de priorité $\mathcal{S}$ basé sur le seul intérêt scientifique des cibles. $\mathcal{S}$ n'est pas figé à ce jour mais nous devons gérer le cas où il entre en conflit avec le S/B . Pour illustrer ce point,

on supposera qu'on souhaite pour des raisons scientifiques faire la photométrie précise d'une étoile donnée (par exemple une étoile chaude) mais qu'elle est contaminée et que son patron empiète sur celui d'une autre étoile, celle-ci brillante avec un fort S/B . Ici encore, tant qu'on n'a pas de correspondance entre le critère $\mathcal{S}$ et le S/B , on ne sait pas en général résoudre le problème de façon automatique. Nous verrons qu'en pratique, s'il existe un lien même ténu entre les deux critères, l'attribution selon $\mathcal{S}\times$ q (la simple multiplication) donne un résultat très satisfaisant. Il permet en outre de disposer du logiciel, sans devoir attendre que $\mathcal{S}$ soit décidé.

$\mathcal{S}$ peut être une classe de magnitudes, une contamination maximale autorisée ou une classe de luminosité qui préside au diamètre de l'étoile. Ces critères ne co-évoluent pas de manière nette avec le S/B . La Fig. 7.2 illustre ce fait dans le cas de la contamination.

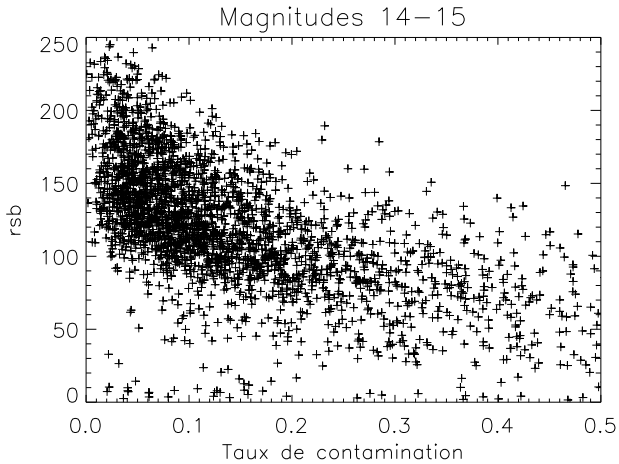


FIG. 7.2 – *Rapport signal à bruit en fonction du taux de contamination pour une gamme de magnitudes donnée. Bien que le lien soit évident le S/B peut varier du simple au double pour une contamination donnée.*

On peut résoudre complètement le problème dans le cas trivial d'une priorité $\mathcal{S}$ et d'un S/B tous deux booléens. Ecrivons la table de vérité du comportement à adopter en cas de collision (Tab. 7.5).

On reconnaît la table de la fonction "ET" logique, la priorité est donc dans ce cas une multiplication Booléenne :

$$P_{\mathrm{Bool}} = \mathcal{S}.q$$

TAB. 7.5 – *Table de vérité à double entrée $\mathcal{S} \times q$. La sortie 1 indique la décision : l'étoile candidate est conservée comme cible.*

$\mathcal{S}$	1	0	1
	0	0	0
		0	1
		rsb	

Cette règle reste intuitivement valable pour la multiplication entre nombres réels, dans le cas de priorités continues, dès lors que les distributions sont fortement piquées autour de 0 et de 1. En effet il est souhaitable que les étoiles de grand intérêt scientifique et de fort S/B soient attribuées en priorité.

Voyons si l'on peut généraliser cette solution à toutes les distributions de priorités. Envisageons pour cela quatre cas réalistes :

H0 : Hypothèse de référence $\mathcal{S} = 1$: la seule priorité est le S/B , c'est le cas actuel ;

H1 : Priorités booléennes : les étoiles contaminées à plus de 10% sont simplement éliminées des cibles,

H2 : Priorités discrètes : on supposera des tranches de magnitudes assorties des priorités scientifiques suivantes :

$$\begin{aligned} m_V \leqslant 14 &\Rightarrow \mathcal{S} = 3/3 \\ m_V \leqslant 15 &\Rightarrow \mathcal{S} = 2/3 \\ m_V \leqslant 16 &\Rightarrow \mathcal{S} = 1/3 \\ m_V > 16 &\Rightarrow \mathcal{S} = 0 \end{aligned}$$

H3 : Priorités continues : La priorité décroît quand le diamètre de l'étoile augmente afin de donner un équivalent des $\Delta F/F$ des transits en termes de rayon des planètes.

Les résultats sont présentés dans la table 7.6. Pour des raisons techniques, H3 n'a pas été testée mais les autres cas sont suffisants pour juger.

Les résultats sont très voisins, quoique légèrement moins bons pour H1 et H2. Les étoiles ayant perdu leur S/B optimal se retrouvent pour la plupart rétrogradées dans la bande des -10% à -50%. Ceci n'a rien de surprenant, en effet (pour H1) on les classe suivant la contamination et on les mesure suivant le S/B . L'effet individuel de H1 sur les étoiles est présenté figure 7.3. Les pics

TAB. 7.6 – *Comparatif des différentes hypothèses H0, H1, H2. Nombre de cibles affectées qui perdent respectivement moins de 1%, 10% et 50% de leur S/B nominal*

Δ S/B	H0	H1	H2	H3
$\leqslant 1\%$	5 753	5 723	5 686	
$\leqslant 10\%$	130	124	137	
$\leqslant 50\%$	114	140	168	
$> 50\%$ ($\equiv$ perdue)	3	13	9	

négatifs et les zéros sont les étoiles ayant perdu du S/B ou ayant cédé leur place. L'élément important est que l'impact de $H \neq H0$ reste cantonné aux étoiles faibles.

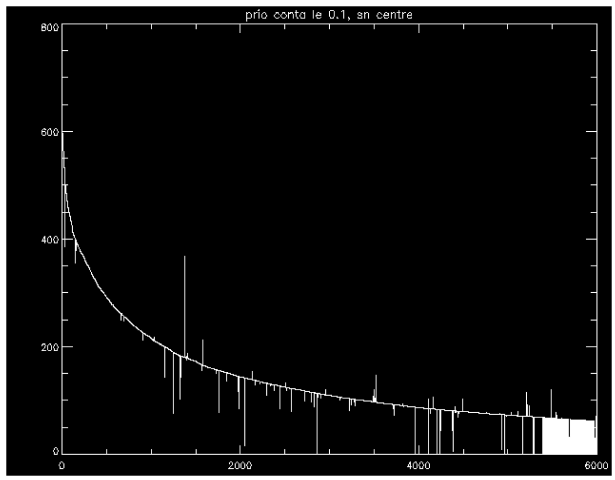


FIG. 7.3 – *Effet individuel de H1. Les étoiles affectées sous H0 sont classées par S/B décroissant. Elles sont ensuite réaffectées suivant H1 et superposées en suivant le même classement.*

On constante l'apparition d'un pic positif pour une étoile (désignons la par 'X') en position 1 400. Ceci peut paraître surprenant en effet, comment une étoile peut-elle doubler son S/B simplement quand on en exclut d'autres plus mineures qu'elle (avec H1) ? L'explication tient au fait que, pour des raisons pratiques nous n'avons pas éliminé complètement les étoiles, mais leur avons donné une priorité négligeable. On voit figure 7.4 que sous H0 le masque de X est bloqué par une étoile 'Y' qui présente un S/B supérieur malgré une contamination dépassant 10%. Sous H1, l'étoile Y se trouve rejetée en fin de classement et la fenêtre de 'X' peut s'étendre. On en voit la contrepartie sous forme d'un pic négatif pour Y visible dans les forts S/B près de l'origine. C'est la conservation d'une priorité

minuscule qui fait que peu d'étoiles sont absentes avec H1, sinon Y aurait été perdue. Le bilan est donc défavorable à H1, ce qui est logique puisque la métrique qui fixe les priorités d'affectation (contamination) est différente de celle utilisée pour la mesure de la performance (S/B).

FIG. 7.4 – *Vue locale de la collision entre l'étoile 1 400 marquée d'un 'X' et une étoile de fort S/B mais contaminée 'Y'. Sous H0 (à gauche) Y bloque logiquement X, mais sous H1 (à droite) Y est rejetée en fin d'affectation.*

Il en va du même genre de scénario pour H2. Les variations brutales d'une priorité discrétisée ne changent rien à l'affectation car le S/B est fortement lié à la magnitude. Les exception sont les rares cas de collision impliquant des étoiles à la frontière de deux classes.

En conclusion une priorité à double critère $\mathcal{S}$ et S/B produit peu d'effet. Pour que celui-ci se fasse sentir il faut : 1) qu'il y ait collision (moins de 20% des cas), 2) que cette collision concerne des étoiles de S/B voisin, ce qui est très rare pour les étoiles fortes qui sont peu nombreuses. 3) même dans ce cas, si les critères sont liés le résultat sera souvent le même qu'avec le seul S/B . *Notre priorité mixte* $P_{Bool} = \mathcal{S}.q$ *répond donc bien à la question et permettra des changements ultérieurs dans le choix des priorités scientifiques sans nécessité de retouche.*

7.5 Données destinées à la détection des transits

Lors de l'attribution, le programme est amené à séparer dans chaque fenêtre les photons de l'étoile de ceux des contaminants. La fraction de PSF $f_j(i)$ de l'étoile i dans la fenêtre j est connue, quelles que soient i et j ; on stocke cette information intermédiaire afin de la rendre disponible pour la détection des transits. En voici deux utilisations éventuelles :

1. **Décontamination**. En cas de chevauchement entre les PSFs d'une cibles 1 et d'une cible 2, une partie du flux de l'étoile 2 est incorporé dans la fenêtre de l'étoile 1. L'activité stellaire de 2 polluera donc l'étoile 1. Pour corriger cette pollution dans la courbe de lumière de la fenêtre 1 (f_1), il faut en retrancher $f_1(2)$. Mais au sol on ne connaîtra pas le flux de l'étoile 2. On sait simplement qu'il domine sa propre courbe de lumière $f(2)$. On utilisera donc l'approximation suivante :

$$s_1 = f(1) - f(2)\frac{f_1(2)}{f_2(2)}$$

où s_1 est le signal corrigé. On pourra donc à l'aide de ces informations atténuer dans une courbe de lumière, la part de contamination imputable aux étoiles des courbes voisines.

2. **Localisation d'une source** : Supposons qu'on détecte un candidat transit dans la fenêtre j. Un problème important est de s'assurer qu'il ne provient pas de l'éclipse de l'une des étoile de fond qui serait binaire. On utilise l'information spectrale fournie par le prisme comme information spatiale de nature à faciliter la discrimination entre un transit sur la cible et une éclipses d'arrière plan. On dispose des $f_j(i)$ de toutes les étoiles i de la fenêtre. Les fenêtres sont dissociées en 3 canaux r, v, b. Considérons un transit h_r, h_v, h_b observé dans les trois courbes correspondantes. Les écarts-type de ces courbes sont $\sigma_r, \sigma_v, \sigma_b$ (hors transit). Plusieurs étoiles de la fenêtre peuvent être suspectées d'avoir donné naissance à h si elles sont binaires à éclipse. Pour chaque étoile i on fabrique l'indicateur :

$$\chi_i^2 = 1/3.\frac{\left(h_r - \tilde{h}.\frac{f_r(i)}{f_r}\right)^2}{\sigma_r^2} + \frac{\left(h_v - \tilde{h}.\frac{f_v(i)}{f_v}\right)^2}{\sigma_v^2} + \frac{\left(h_b - \tilde{h}.\frac{f_b(i)}{f_b}\right)^2}{\sigma_b^2}$$

$\tilde{h}$ est le $\Delta F/F$ inconnu à l'origine de l'événement observé. f_r, f_v, f_b sont les flux totaux des bandes rouge, verte et bleue. Un χ_i^2 faible indique une forte probabilité pour l'étoile i d'être à l'origine de l'événement observé. On range χ_i^2 calculé pour toutes les étoiles de la fenêtre dans une table (voir Fig. 7.5).

La probabilité pour une étoile donnée de mimer l'événement observé est maximale quand le pic est haut. Plus le pic est pointu, meilleure est l'estimation du responsable et son $\Delta F/F$. La cible est un candidat comme un autre. La gamme des $\Delta F/F$ commence à 10^{-4} pour un transit sur la cible, et augmente jusqu'à 50%, pour l' éclipse de plein fouet d'une étoile de fond. La table peut servir à déterminer l'origine d'une observation si le pic

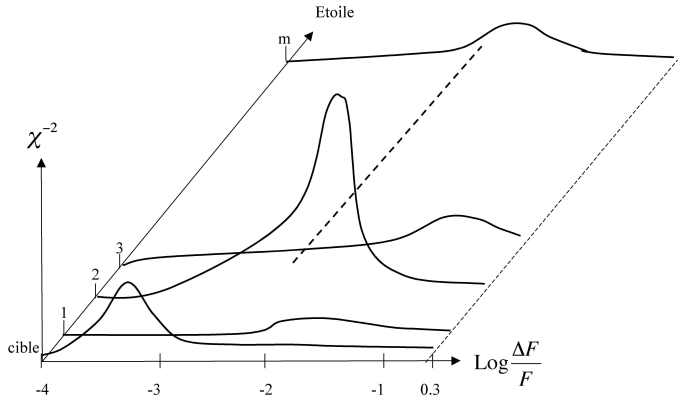


FIG. 7.5 – *Inverses du χ^2 pour chaque étoile de la fenêtre et pour différents niveau d'occultation $\tilde{h}$. L'un des point est à l'origine de l'événement observé.*

est haut, ou simplement à éliminer les étoiles et amplitudes étrangères au phénomène quand aucun pic ne se distingue. Dans ce dernier cas on préfère l'indicateur

$$\chi'^2 = \max_{q=\{r,v,b\}} \frac{\left(h_q - \tilde{h}.\frac{f_q}{f}\right)^2}{\sigma_q^2}$$

qui présentera des creux plus marqués dès qu'une couleur sera incompatible avec la couleur de l'observé.

Pour finir, les PSFs ne se limitent pas aux frontières d'un seule fenêtre. On pourrait aller chercher de l'information dans les fenêtres voisines en ne considérant plus des couleurs mais des cannaux indifférenciés. Pour l'étoile i :

$$\chi_i{}^2 = 1/n \sum_j \frac{\left(h_j - \tilde{h}.\frac{f_j(i)}{f_j}\right)^2}{\sigma_j^2}$$

Où j désigne tout canal ayant une source de photons commune avec ceux où le transit à été observé. Cette technique n'a pas encore été quantifiée à l'heure actuelle.

Afin de préciser la magnitude des étoiles de fond dont les éclipses seraient gênantes, on a simulé figure 7.6 des éclipses de plein fouet dans chacune des fenêtres d'un champ du centre galactique. Les points sont les plus petits transits détectables sur la cible, les lignes obliques sont les profondeurs de transit que mimerait une étoile binaire située prés de la cible. On voit que les éclipses d'étoiles de fond plus brillantes que $m_V = 20$ seront toujours

visibles. Celles d'étoiles $m_V = 21$ deviennent gênantes pour des cibles plus faibles que $m_V = 13.5$, et les étoiles de fond $m_V \geqslant 22$ se sont jamais détectables.

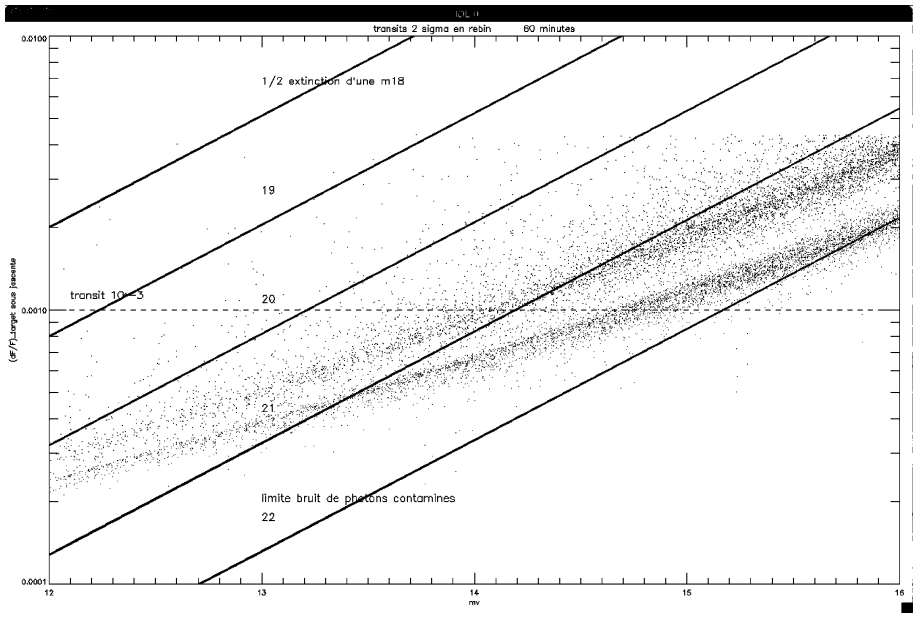


FIG. 7.6 – *Détectabilité des étoiles de fond binaires à éclipses. L'affectation à eu lieu pour un champ de l'anticentre galactique. Les étoiles sont classées par magnitude. Chaque point du nuage supérieur est la taille du plus petit transit détectable, en lui supposant une amplitude $\Delta F \geqslant 2\sigma$ sur une heure. Ces valeurs tiennent compte de la contamination, des bruits de photon, des bruits électroniques, du jitter et de la respiration. Le nuage inférieur ne considère que le bruit photonique de la cible et sa contamination. Les lignes obliques sont les transits équivalents pour des éclipses de plein fouet d'étoiles de fond située dans le même pixel que la cible. Le trait pointillé horizontal est le transit à 10^{-3}.*

7.6 Conclusion et perspectives pour le fenêtrage

Les procédures de réduction et d'affectation sont maîtrisées et donnent de bons résultats avec les PSFs théoriques. En aval de toutes ces procédures, les étoiles perdent rarement plus de quelques pour cent du S/B qu'elles auraient si elle étaient des cibles isolées munies de leur masque optimal.

Ces résultats sont obtenus pour des PSFs théoriques connues avec la précision nécessaire. L'étape suivante consiste à savoir préserver ces performances avec les

PSFs réelles de Corot en vol. Ces PSFs vont différer des PSFs théoriques car elles découleront de mesures effectuées à partir d'une image prise en vol. Sur une telle image, on ne peut pas séparer avec précision les cibles de leur contamination. Il s'ensuit que la précision des PSFs déduites sera très limitée ; on parle de plus de 20% d'erreur à l'endroit du pic de la PSF. Il faudra donc être capable de calculer des masques optimaux, de savoir les réduire et les affecter en dépit d'un S/B peu précis.

Deux éléments devraient cependant jouer un rôle dans la résolution de ce problème :

- Tout d'abord ce sont les PSFs des étoiles faibles qui seront connues avec le moins de précision, or on a vu que ces étoiles étaient justement les plus tolérantes aux masques peu adaptés ;
- Ensuite, la connaissance précise du flux total dans chaque pixel permettra tout de même de connaître deux des trois termes de bruits intervenant dans le S/B : le bruit photonique et le jitter. Il ne restera qu'un nombre limité d'hypothèses à faire quant à la valeur du flux de l'étoile.

Le nombre d'hypothèses à faire s'en trouve donc limité.

Deuxième partie

Détection des transits dans les courbes de lumière

Chapitre 8

Énoncé des contraintes.

Le but recherché est de détecter les plus petits transits possibles dans 12 000 courbes de lumière contenant plusieurs types de bruit tout en évitant au maximum les fausses détections.

La détection doit être adaptée aux bruits de diverses natures, aléatoire (bruit de photons), stellaire (activité), environnemental (résidus de lumière diffusée) ou technique (impacts de protons, jitter ou respiration résiduelle).

Il faut également être en mesure de distinguer les transits détectés des événements d'une autre nature, notamment les éclipses d'étoiles doubles ou triples situées en arrière plan, les taches sombres sur une étoile en rotation, etc...

En dépit des différents types de bruit, il faut parvenir à estimer avec précision l'amplitude $\Delta f/f$, la durée et la période des transits,

On pourra également exploiter l'information des trois bandes colorées r, v, b qui est présente dans environ 10 000 des courbes,

L'algorithme de détection doit également s'accommoder des données manquantes, dues notamment aux impacts de rayons cosmiques sur le CCD lors du survol de l'anomalie magnétique de l'Atlantique Sud (SAA), au dépointage lors de la rotation des panneaux solaires, aux calibrations ou encore aux opérations de maintenance.

On essaiera aussi de travailler au fur et à mesure de l'arrivée des données afin de mettre les événements détectés à disposition le plus tôt possible.

Quand j'ai débuté mon travail sur la détection, les méthodes utilisées n'avaient pas encore exploité la multiplicité des courbes de lumière, qui est un

avantage spécifique de Corot . J'ai montré dans le cadre d'un test en aveugle que l'on pouvait améliorer les capacités de détection par recoupement des différentes courbes. La statistique permet de reconnaître parmi le bruit des variations qui sont en fait déterministes. J'ai utilisé un algorithme de détection simple dérivant du "filtrage adapté", pour tester plusieurs façons de déceler et d'exploiter ces déterminismes communs. En procédant par pondération des artefacts synchrones, on fait émerger du continuum 80% d'événements supplémentaires par rapport à la détection simple.

En dépit de la simplicité du détecteur utilisé, le résultat s'inscrit dans la moyenne de ceux des autres équipes participantes au test en aveugle attestant de l'intérêt d'une approche collective du débruitage.

8.1 Introduction

Depuis l'espace, Corot transmettra les données de 60 000 étoiles mesurées toutes les 8,5 min dans des conditions identiques et pendant 150 jours. Corot est dimensionné pour conserver une précision $\sigma = 7.10^{-4}$ jusqu'à la magnitude 15.5. Avant de pouvoir être exploitées par les algorithmes de détection, les courbes de lumière sont prétraitées par un pipeline de données qui les calibre et les débarrasse des effets instrumentaux et environnementaux connus tel le jitter, la lumière diffusée, les impacts de particules ionisantes, etc ...Les algorithmes de détection interviennent après cette étape. Pour détecter les planètes les plus petites, ces algorithmes doivent exploiter au maximum l'information disponible sur le signal et sur le bruit.

Le problème de la détection d'un signal donné dans un environnement bruité est un des problèmes de base du traitement du signal. La spécificité de la détection des transits planétaires est qu'on cherche à la fois une forme et une période. On peut chercher les transits dans l'espace direct en comparant la courbe temporelle avec une référence de transit, on peut aussi prospecter dans l'espace des fréquences car ce sont les seuls signaux ponctuels parfaitement périodiques. La difficulté est de combiner ces deux indications en un critère de détection unique.

8.2 Les méthodes de détection

Plusieurs méthodes sont proposées pour la détection. Certaines recherchent un gabarit dans le signal temporel, d'autres combinent forme et fréquence.

Mais aucune méthode efficace n'est basée sur l'analyse de Fourier. En effet, les transits sont des événements brefs, leur faible énergie se répand donc dans une large gamme de fréquences dont aucune n'a tendance à émerger du bruit. Comme l'explique Bordé (2003) seul $\simeq 7\%$ de l'énergie d'un transit de durée analogue à celui de HD 209458 b se trouve à la fréquence du fondamental.

La méthode de corrélation est la méthode la plus classique. Elle est optimale au sens du maximum de vraisemblance pour un signal en présence de bruit Gaussien additif non corrélé. Il s'agit de calculer la valeur de la corrélation :

$$\rho = \sum_t s(t)x(\Delta f, d, T, t+\phi)$$

entre le signal observé $s(t)$ où t est un temps discret, et un gabarit multi-transit de référence de même longueur à quatre paramètres $x(\Delta f, d, T, \phi)$, Δf étant la profondeur, d la durée, T la période et ϕ la phase à tester. Il s'agit d'essayer toutes les valeurs possibles de $\Delta f, d, T, \phi$, le jeu le plus probable étant celui qui maximise ρ. On voit que le nombre des combinaisons est très élevé, mais on peut le réduire par les considérations suivantes :

- Δf n'a pas besoin de varier, la multiplication par une constante ne change pas les valeurs optimales pour les autres paramètres, on choisit donc $\Delta f = 1$,
- On calcule la corrélation pour toutes les valeurs de ϕ en une seule opération à l'aide de la transformée de Fourier discrète (TFD). En effet, la TFD de la corrélation vue comme une fonction de la variable ϕ peut s'écrire :

$$\mathrm{T}FD\left[\sum_t s(t), x(t+\phi)\right] = S^*.X \tag{8.1}$$

  où S, resp. X sont les transformées de s, resp. x, et * désigne le complexe conjugué. La phase optimale $\tilde{\phi}$ est donc la phase qui maximise :

$$\mathrm{T}FD^{-1}(S^*.X).$$

- La quantité de calculs se trouve réduite pour les autres paramètres ce qui permet d'utiliser un maillage suffisamment serré pour d et T, après avoir contrôlé que les pics de détection tolèrent la largeur de maille souhaitée.

Cette méthode n'est optimale que dans le cas Gaussien, mais un filtrage préalable des courbes de lumière permet d'atténuer les bruits trop éloignés de cette hypothèse. Cette méthode est fréquemment nommée "filtre adapté", duquel elle n'est en fait qu'un cas particulier (Defaÿ 2001).

Dans le cas où chaque point possède son propre écart-type $\sigma(t)$, le filtre adapté s'écrit :

$$\rho = \sum_t \frac{s(t)x(t)}{\sigma^2(t)}.$$

Le cas le plus général du filtre adapté est optimal pour les bruits corrélés. Une littérature abondante s'y rapporte (voir par ex. Kay (1998)). Il exploite les corrélations internes du bruit et s'écrit :

$$\rho = \frac{x^t R^{-1} s}{\sqrt{s^t R^{-1} s}}$$

où R^{-1} est l'inverse de la matrice d'autocorrélation du bruit (supposé centré). On suppose pour simplifier qu'on peut se ramener au cas où R est inversible. Pour percevoir le rôle de R, plaçons-nous dans la base orthogonale de ses vecteurs propres (on a vu au §6.4.3 que R était symétrique, définie et positive). Dans ce cas R^{-1} est la matrice

$$R^{-1} = \begin{pmatrix} \sigma_1^{-2} & 0 & \cdots & 0 \\ 0 & \sigma_2^{-2} & & \vdots \\ \vdots & & \ddots & \vdots \\ \vdots & \vdots & & \sigma_n^{-2} \end{pmatrix}$$

où σ_i est à présent l'écart-type suivant le $i^{ème}$ vecteur propre. L'expression de ρ équivaut alors à :

$$\rho = \frac{\sum \frac{x_i s_i}{\sigma_i^2}}{\sqrt{\sum \frac{s_i^2}{\sigma_i^2}}}$$

les x_i, resp. s_i étant les coordonnées de x, resp. s dans cette base. C'est-à-dire que les entités comparées ne sont plus les points de mesure (qui sont des quantités interdépendantes), mais leurs combinaisons linéaires (qui elles, sont indépendantes). Dans notre cas, la difficulté est de choisir une matrice R qui soit bien représentative du bruit. On peut songer à s'aider d'une classification des courbes de lumière par types spectraux ou par similitudes du bruit.

Pour les planètes tournant autour d'étoiles doubles analogues à *CM draconi*, les conformations de gabarits décrivant les trois corps sont très nombreuses. Jenkins et al. (1996) a étudié dans ce cas l'usage du filtre adapté.

De leur côté, Deeg et al. (1998) ont pris en compte la variabilité atmosphérique suivant la quantité d'atmosphère traversée en construisant un profil de

pondération estimé à partir de nombreuses observations.

Suivant une approche différente, la méthode Bayésienne de Defaÿ et al. (2001) estime la forme du transit à partir de 7 harmoniques. Le fondamental le plus probable est obtenu par le maximum de vraisemblance. Aigrain & Favata (2002) utilisent également une approche Bayésienne du problème.

Une autre classe de méthodes combinent de manière constructive le facteur de forme et de période. Le "Box Least Square" (BLS) proposé par Kovács et al. (2002) commence par replier la courbe de lumière sur elle même à une période arbitraire T puis calcule la ressemblance entre la courbe repliée et un transit de référence choisi rectangulaire. La nouvelle courbe $x(t), i = 1, 2, \ldots, n$ est constituée de super-points regroupant les points initiaux en coïncidence. Un super-point x_i est affecté du poids :

$$w(t) = \left[\frac{\sigma^2(t)}{\sum_{j=1}^{n} \sigma^2(j)} \right]^{-1} .$$

où $\sigma(t)$ est l'écart-type des points constitutifs du super-point t. L'utilisation d'un gabarit rectangulaire permet de calculer $\tilde{L}$ la profondeur optimale du transit, directement à partir des données ce qui fige l'un des paramètres libres, les autres étant d, T et ϕ. $\tilde{L}$ est calculée avec les points internes au transit par :

$$\tilde{L} = \frac{\sum w_i x_i}{\sum w_i}$$

On reconnaît l'expression de leur barycentre. L'algorithme explore l'espace des paramètres après avoir déterminé le maillage optimal. La ressemblance avec le gabarit est évaluée par un test de χ^2. Cette méthode à été employée avec succès dans le cadre de l'expérience OGLE. Tingley (2003b) note une correspondance de formulation entre le BLS à base de χ^2 et le filtre adapté.

Pour leur part, Aigrain & Irwin (2004) utilisent une approche similaire avec un gabarit en créneau dont ils calculent le χ^2 aux courbes repliées. Ils effectuent un lissage préalable destiné à uniformiser le niveau moyen local de la courbe avant repliement et utilisent une pondération adéquate des points dont la mesure ne tombe ni entièrement dans le transit, ni entièrement en dehors.

D'une manière générale, quand les transits deviennent faibles il faut pousser la sensibilité en modifiant un seuil, pour continuer de les détecter. Mais cette opération s'accompagne d'une croissance rapide des fausses détections. Les méthodes évoquées sont d'autant plus efficaces qu'elles opèrent sur des courbes débruitées. Par exemple les variations à long terme dues à l'activité stellaire

peuvent avoir une amplitude supérieure à celle des transits, risquant d'oblitérer les données dans les courbes repliées.

8.3 Test en aveugle

Afin de préciser les idées quant à la détectivité de Corot , des méthodes existantes adaptées pour Corot ou développées ad hoc par cinq équipes ont été comparées dans le cadre d'un test en aveugle. Mis en place par Moutou et al. (2005), ce test est basé sur 1 000 courbes simulées semblables à celles que produira Corot . Des événements de nombre et de nature non-révélés ont été ajoutés dans certaines. Il s'agit de comparer la capacité des différentes méthodes à détecter de petits transits en évitant les fausses détections, et d'évaluer l'impact de la variabilité stellaire et des étoiles de fond binaires à éclipses.

8.3.1 Les données de simulation

Les courbes ont été produites à l'aide d'un simulateur d'instrument. Elles incluent la micro-variabilité stellaire, et quelques dizaines de transits planétaires et stellaires. Les algorithmes se révèlent d'efficacité différentes, tant pour détecter les transits que pour éviter les fausses détections. Le test souligne également l'importance du "débruitage" des courbes. Ce sont les algorithmes à base de repliement temporel qui se comportent le mieux. Les fausses détections sont différentes entre les méthodes mais pas les vraies, ce qui favorise leur discrimination.

Voici une liste des bruits inclus dans la simulation et qui n'ont pas été divulgués durant le test, ainsi que les techniques possibles pour s'en affranchir :

- Le bruit de photon d'écart-type $\sqrt{F}$. Il suit une loi quasi-Gaussienne. Moyenner n échantillons réduit l'écart-type d'un facteur $\sqrt{n}$ et équivaut à un temps de pose n fois plus long. Mais cette opération peut déformer le transit ou y incorporer des points anormaux.

- La lumière diffusée par la Terre. Dans cet exercice, elle varie le long de l'orbite. On en simule une compensation imparfaite par le pipeline à l'aide d'un profil temporel d'amplitude $\simeq$0.5 e^-/pixel/seconde pondéré d'un coefficient résiduel aléatoire, positif ou négatif. Ce résidu peut engendrer une amplitude maximale de $\sigma/F \simeq 1\%$ pour une étoile $m_V = 14$ et un masque de 80 pixels, c'est-à-dire supérieure à l'ordre de certains transits.

Les résidus de correction proviennent de l'utilisation de modèles prédictifs imparfaits. On peut envisager d'affiner le modèle à l'aide des données sur le long terme. Il en va de même pour la correction du jitter, mais l'opération est plus complexe car elle dépend d'une PSF en deux dimensions.

– La variabilité stellaire. Elle est injectée de deux façons, soit à partir de son spectre (Aigrain et al. 2003), soit à l'aide de modèles de zones actives Lanza et al. (2003). Moins nombreux que les courbes de lumière, les profils de variabilité sont dilatés et déphasés arbitrairement pour éviter leur répétition.

  Elle peut dans une certaine mesure être modélisée pour en dégager un profil temporel déterministe comme l'ont fait Lanza et al. (2003) avec succès pour le Soleil. On peut aussi s'attacher à ses caractéristiques fréquentielles déterminées par Aigrain et al. (2003), enfin on peut la caractériser de manière statistique.

– Les impacts de protons. Ils sont très abondants pendant $\simeq$ 15 min toutes les 1.7 heures, lors de la traversée de l'anomalie magnétique. Leur détection par l'électronique de bord est simulée par l'absence des mesures correspondantes,

  Les trous de données qu'ils causent ne sont pas des bruits, mais compliquent la détection. On peut les interpoler ou utiliser des méthodes de détection qui les tolèrent. Certains rayons cosmiques de basse énergie échappent à la détection basée sur un seuil. On peut tout de même déceler l'origine non lumineuse d'un pic en comparant les trois courbes de lumière d'une fenêtre colorée.

– On remarque la présence de points fortement excentrés. Ils sont de nature à induire un biais si on les combine à d'autres points par une opération linéaire. Ignorant leur origine, on ne peut pas les rectifier à l'aide d'un modèle. On se contente d'en réduire le nombre par des combinaisons à base de seuil ou de médiane. La difficulté est qu'on ne sait pas à partir de quel écart un point doit être considéré comme anormal. En fait ils étaient simplement dus à la combinaison des autres bruits listés.

– Les bruits non listés sont mineurs. Il s'agit de la non-uniformité de réponse du CCD (1%), du bruit de lecture ($10e^-$/pixel/seconde) et de la lumière zodiacale ($12e^-$/pixel/seconde). Ils sont uniformes ou constants, onles suppose corrigés et on n'en conserve que le bruit de photons.

– Le jitter ne fait pas partie de la simulation,

Pour finir, des transits de $1.6R_{\oplus}$ à $1.3R_J$ ont été inclus dans certaines courbes, ainsi que différents autres phénomènes stellaires, étoiles triples, transits stellaires, binaires rapprochées. Ni leur nature, ni leur nombre et paramètres ou ceux des bruits n'étaient connus.

La figure 8.1 montre deux exemples de courbes de lumière ainsi simulées, l'une avec et l'autre sans transit. A long terme les courbes sont dominées par la variabilité stellaire basse fréquence et le fort écart-type. A une échelle plus brève d'une dizaine d'orbites, on remarquerait surtout les trous, la lumière diffusée périodique et les points anormaux. L'ensemble des bruits est souvent d'un ordre égal ou supérieur à celui des transits potentiels.

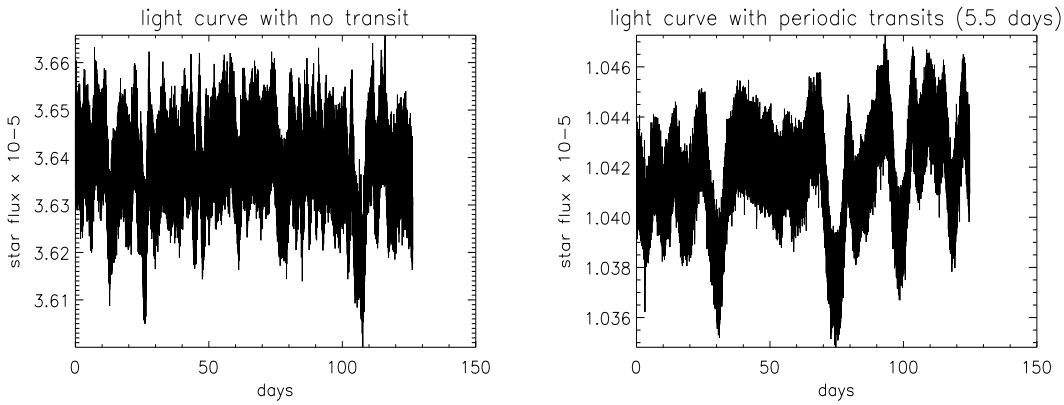


FIG. 8.1 – *Courbes de lumière brute n°1 (gauche) et n°34 (droite). Seule celle de droite contient des transits, qui sont parmi les plus visibles.*

8.3.2 Les méthodes utilisées

Toutes les méthodes utilisées pour le test en aveugle sauf celle proposée, ont en commun une phase de filtrage qui précède la détection proprement dite. Le filtrage sert à débarrasser les courbes de la variabilité stellaire et à les nettoyer des bruits les plus évidents, pour permettre à l'algorithme de détection de se focaliser sur les transits.

Les anomalies les plus visibles sont la lumière diffusée et la présence de données manquantes. Comme le test est organisé en aveugle, l'origine des bruits

inclus n'était pas communiquée. Ceci associé au fait que les équipes travaillaient séparément à conduit à une grande diversité de techniques de filtrage, décrites dans Moutou et al. (2005). On trouve principalement :

La normalisation : Les courbes peuvent être ramenées à l'échelle des transits $\Delta f/F$ en les divisant par F.

Les données manquantes : – Elles sont parfois interpolées linéairement,

– Les trous peuvent aussi être comblés par des procédures de lissage des points anormaux, filtre médian ou moyenne glissante,

– Dans certains cas, les courbes conservent leurs trous et ce sont les traitements qui s'adaptent.

La lumière diffusée : – Le plus fréquent est le filtre passe bas car la lumière diffusée se trouve aux fréquences $f \geqslant f_{\mathtt{orb}}$. Les filtres utilisés peuvent être une moyenne glissante, ou un lissage médian itératif alternant avec la suppression des points éloignés de plus de 3σ ce qui supprime du même coup les points anormaux.

– Comme l'origine de nombreux bruits est un mouvement circulaire, l'équipe 3 a ajusté la somme d'harmoniques sinusoïdaux qui explique au mieux le signal observé. La fréquence fondamentale de la lumière diffusée est mesurée à $T = 1.7$ h. Leur technique s'adresse en même temps à la variabilité stellaire en utilisant un autre fondamental à $T = 300$ j. Les harmoniques de ce dernier sont choisis pour ne pas englober de transit dans l'ajustement. Le système d'équation surdéterminé est résolu à l'aide d'une décomposition en valeurs singulières (Press et al. 1997) qui minimise la distance entre la solution retenue et l'observé. Les données manquantes ne font simplement pas partie du système d'équation.

– La courbe est repliée à la période $T = 1.7$ h, lissée pour obtenir le profil périodique de lumière diffusée (constant dans la simulation mais pas dans la réalité), puis soustraite,

– On trouve enfin un procédé de sous-échantillonnage suivi d'un ré-interpolation en lissant indépendamment les parties inférieures et supérieures de la courbe considérée comme une image à deux dimensions.

La variabilité stellaire : Elle est approximée globalement par un filtre

passe-bas, ou bloc par bloc à l'aide d'une droite où d'un polynôme dont on minimise la distance à la courbe au sens des moindre carrés. Le modèle réalisé est ensuite soustrait. Pour limiter le risque d'ajuster les transits en même temps que la variabilité, la taille des blocs (12h-72h) est choisie supérieure à la durée supposée des transits.

La figure 8.2 montre un exemple de processus de débruitage. En haut, la courbe est brute ; au centre, les trous sont comblés et un filtre passe-bas à été appliqué. En bas la variabilité stellaire à été modélisée puis soustraite. On voit beaucoup mieux apparaître les transits.

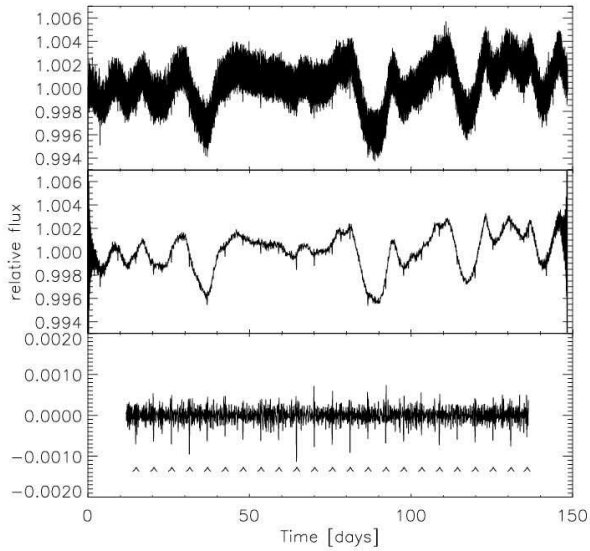


FIG. 8.2 – *Haut : courbe brute, les trous ne sont pas visibles à cette échelle. Centre : courbe filtrée passe-bas ($f_c \sim 1.5h$). Bas la variabilité modélisée est soustraite.*

Les méthodes de détection employées reposent sur le filtre adapté ou le repliements des courbes à des périodes d'essais. Pour plus de détails se référer à l'article de Moutou et al. (2005) situé en annexe.

8.3.3 Traitement des données manquantes

Des méthodes différentes ont été utilisées pour gérer les données manquantes. Certains pré-traitements comblent les trous en les interpolant, d'autres non.

Il est à l'évidence plus confortable d'interpoler les données à l'aide par exemple d'une fonction linéaire, médiane ou cubique de Spline (Press et al.

1997). Mais le comblement des trous risque de diluer l'information disponible dans de l'information factice. En effet, supposons l'interpolation linéaire de deux point p_1 et p_3 entourant un trou. L'utilisation du point interpolé $p_2 = (p_1 + p_3)/2$ rendra la courbe plus "lisse", mais ne fera que donner un poids exagéré à p_1 et p_3. Une détection sur une courbe interpolée équivaut à chiffrer la ressemblance d'un gabarit avec des points qui n'existent pas. En toute rigueur la détection doit procéder à l'inverse : appauvrir le gabarit des points manquants dans la courbe et faire la détection avec ce qu'il en reste.

Mais d'un autre côté un échantillonnage régulier et sans trous permet l'accès à des méthodes éventuellement plus efficaces. Le gain pour la détection peut se révéler supérieur au biais introduit. Il n'y a pas de règle universelle pour choisir, cela dépend de la taille des trous, du bruit et de la redondance des données. Mais si on choisit d'interpoler, la démarche logique est de confirmer après coup la robustesse des détections en faisant varier les points interpolés.

Dans ce travail nous avons préféré ne pas interpoler car :
- Le nombre de données manquantes peut atteindre 15%,
- ces données sont localement denses, regroupées par intervalles continus de 17 à 25 min pouvant perturber des fréquences de $\sim$ 45 min qui sont compatibles avec celles présentes dans les transits.

Chapitre 9

La méthode proposée

La méthode que nous proposons montre que l'on peut améliorer la détection par une caractérisation collective des courbes de lumière d'un même champ. Elle tire parti du fait que les courbes de lumière de Corot sont nombreuses, acquises simultanément et dans des conditions identiques.

Dans un premier temps, la détection directe a permi d'identifier 12 événements. Ensuite l'élimination des caractéristiques communes aux 1 000 courbes a mis à jour 10 événements supplémentaires. Notre stratégie comporte deux spécificités : l'algorithme de détection est appliqué en amont du filtrage d'une part, et de l'autre on utilise l'information disséminée dans l'ensemble des courbes pour pondérer les artefacts ayant survécu à la détection. Cette méthode est celle utilisée par l'équipe numéro 1 de l'article de Moutou et al. (2005) en annexe où elle est décrite à côté des autres méthodes. Elle à également été détaillée dans une publication à part jointe à la fin de ce chapitre.

Le processus commence par l'application d'un détecteur simple dérivant du filtrage adapté. Il s'agit du coefficient de corrélation statistique entre chaque section de la courbe et un transit de référence. Ce détecteur s'adapte au bruit local et le nombre de paramètres des transits de référence à essayer (aussi appelés "gabarits") est réduit. Il produit des courbes de détection temporelles, une par étoile, dont les pics indiquent la probabilité des transits.

La deuxième étape consiste à supprimer les tendances communes parmi les courbes de détection. L'identification de ces tendances se fait au moyen d'une analyse par composantes principales (Press et al. 1997). Seuls sont conservés les pics de détection qui se démarquent significativement de la tendance commune.

La dernière étape est le contrôle de la périodicité et de la nature des événements détectés. Le bruit est moyenné par repliement des courbes brutes

après recentrage individuel des événements. L'examen séparé des occurrences d'indice paires et impaires aide à distinguer les éclipses dissymétriques des étoiles de fond binaires.

Cette chronologie contraste avec celles des autres méthodes du test en aveugle parce que la détection y précède tout filtrage pour rehausser préalablement le contraste des transits recherchés.

9.1 La détection

9.1.1 Définition du débruitage et de la détection

Le *filtrage* et la *détection* sont en fait des notions proches. Par exemple la technique connue sous le nom de "filtre adapté" est en fait une méthode de détection. Afin d'éviter les confusions voici les définitions préalables que nous emploierons par la suite :

Le débruitage consiste à utiliser des algorithmes pour réduire le niveau du bruit d'une courbe de lumière dans le domaine temporel ou fréquentiel. Le résultat du débruitage est une nouvelle courbe de lumière plus douce. Seuls certains bruits sont bien connus comme le bruit de Poisson, par conséquent les procédures de filtrage ne peuvent jamais être parfaitement adaptées. On utilisera indifféremment les termes de débruitage ou de filtrage.

La détection consiste à repérer dans une courbe de lumière les motifs qui ressemblent le plus aux transits. Le résultat de la détection est une surface multidimensionnelle $\rho(\Delta f/f, d, T, \phi)$ paramétrée par l'amplitude $\Delta f/f$, de durée d, la période T et la phase ϕ des candidats transit. Le détecteur que nous employons se limite à produire une courbe de corrélation temporelle $\rho(t)$ de même longueur que la courbe de lumière, qui présente des pics aux instants où un transit est probable.

Dans ce qui suit, on ne parlera que de "point" ou d'"échantillon", le terme "pixel" étant réservé aux images pour éviter toute confusion.

9.1.2 Justification du séquencement inversé

Le signe distinctif de notre méthode est que l'ordre des opérations est inversé par rapport aux autres méthodes. Nous préférons appliquer la détection sur les courbes préalablement au débruitage pour plusieurs raisons :

– L'information la plus précise doit être utilisée la première pour permettre que le rehaussement du contraste soit maximal. Dans le cas de Corot , il s'agit de la forme des transits, de laquelle on possède une description analytique.

– L'énergie des transits est concentrée en temps, ce qui lui donne plus de chances de se démarquer du bruit local. Tout filtrage, temporel ou fréquentiel, implique le mélange d'un point avec ses voisins, diluant le transit avec les échantillons environnants. Pour le vérifier, nous avons appliqué la détection sur une courbe, avant et après filtrage de la lumière diffusée. Le filtrage utilise une moyenne glissante à la période orbitale. Ce filtrage est très efficace contre la lumière diffusée tout en ayant des effets minimes sur les composantes plus lentes. Nous avons utilisé pour cette détection aussi bien un gabarit intact que le filtré de ce gabarit par la même transformation que la courbe. Le résultat pour le gabarit intact est présenté figure 9.1. Les pics de détection sont plus importants dans le cas de la courbe filtrée, mais proportionnellement entourés de plus haut pics de bruit. Au final le S/B de détection, défini par :

$$\text{S/B} = \frac{\rho(t)}{\sigma}$$

où $\rho(t)$ est la hauteur des pics de détection et σ l'écart-type de la courbe de détection, est dégradé par le filtrage. Seule la détection appliquée sur la courbe brute permet de distinguer qu'en réalité deux transits distincts alternés sont présents, signe d'une étoile de fond binaire.

La raison de ce phénomène est que le filtrage dilue le signal de transit avec d'autres points, et tous les signaux estompés ont tendance à se ressembler. Dans le domaine fréquentiel, le filtrage coupe les hautes fréquences du gabarit, dont on a vu qu'elles emportaient une part importante de l'énergie.

– A mesure qu'on veut le rendre plus efficace, les composantes touchées par le débruitage s'approchent de celles du transit et le dégradent de plus en plus. C'est d'autant plus vrai que certains bruits étant peu connus, fixer une limite comporte une part d'arbitraire.

– Une autre raison qui nous incite à ne pas commencer par le débruitage est qu'il est inutile d'éliminer des courbes initiales les composantes qui le seront de toute façon par la détection.

– La dernière raison est que le détecteur choisi est insensible aux données manquantes. Leur seul effet est d'augmenter légèrement le bruit de détection aux dates des trous.

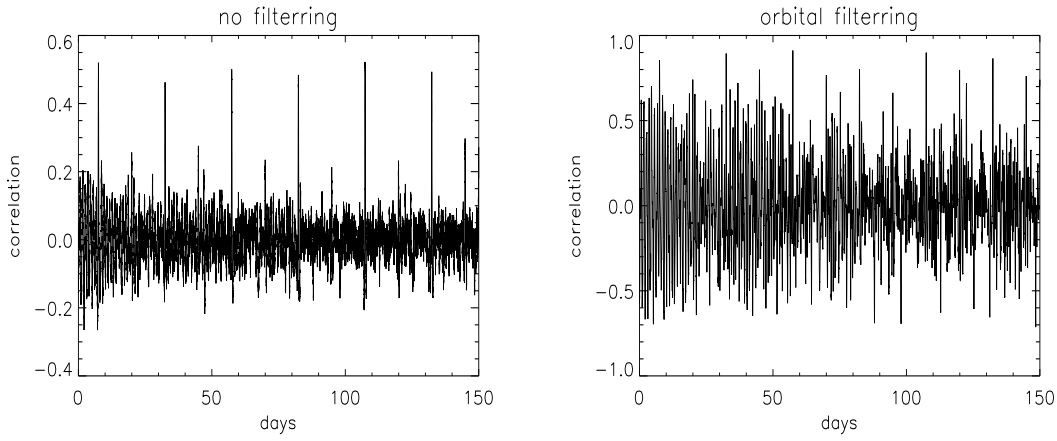


FIG. 9.1 – *1) Détection de transit sur une courbe de lumière brute et 2) après filtrage ciblé de la lumière diffusée. Les pics de détection sont plus hauts après filtrage, mais le S/B de détection a chuté.*

En complément, tout ajout d'un flux correctif distord partiellement le transit embarqué. En effet, ce flux ne contient aucun photon de l'étoile, alors que le flux disparu en contenait. Un transit s'en trouvera donc affecté. En toute rigueur il faut modifier le gabarit en conséquence, mais cet effet reste minime. De plus cet effet n'existe pas pour les corrections portant sur un excès temporaire de flux qui lui, est dû à une contamination passagère et peut être retiré.

9.1.3 Détecteur utilisé

Il faut nous doter d'un détecteur simple et efficace. D'après les essais de Tingley (2003b) (qui ne sont toutefois pas faits avec des courbes de Corot), le filtre adapté et les méthodes à base de corrélation sont celles qui donnent le meilleur résultat. Compte tenu de la forme du bruit nous avons choisi un détecteur de cette famille qui se montre adaptatif au bruit local : le coefficient de corrélation statistique. On se donne un transit de référence x de longueur n (impaire) et une courbe de lumière $s(t)$, la courbe de détection temporelle vaut :

$$\rho(t) = \frac{\overline{x.s(t)} - \overline{x}.\overline{s(t)}}{\sigma_x \sigma_s(t)} \tag{9.1}$$

où la moyenne $\overline{s(t)}$ et l'écart-type $\sigma_s(t)$ portent sur le segment $[t - \frac{1}{2}(n-1), t + \frac{1}{2}(n-1)]$ de s, de longueur n centrée sur t. σ_x est l'écart-type estimé par $\sqrt{\overline{x^2} - \overline{x}^2}$. Le temps t est en réalité un temps discret, c'est-à-dire un

numéro d'échantillon.

Ce détecteur à plusieurs avantages :

- Il s'adapte au bruit et à l'offset local de la courbe à l'instant t car $\overline{s(t)}$ et $\sigma_s(t)$ sont recalculés à chaque position, ce qui n'est pas le cas avec la corrélation qui utilise la transformée de Fourier (cf. 8.1),
- Le détecteur produit une courbe temporelle, ce qui réduit à deux le choix des paramètres initiaux $(\Delta F, d, T, \phi)$ du gabarit,
- Ce détecteur est sensible à la forme, mais pas au facteur d'échelle en ordonnée. Le choix du gabarit se résume à un paramètre unique : d,
- La présence d'un trou n'altère pas significativement la détection. On peut voir ρ comme la mesure du nuage de points statistiques dont l'abscisse est le flux sur la courbe et l'ordonnée le flux sur le gabarit au même instant. La présence des trous se traduit par 15% de points en moins sans changer l'épaisseur du nuage,
- Le résultat est normalisé entre -1 (anticorrélation) et 1 (corrélation parfaite en l'absence de bruit) ce qui permet de comparer les courbes de détection $\rho_i(t)$ entre elles,
- La recherche de la période T et de la phase ϕ se limite aux pics de détection qui dépassent un seuil arbitraire, typiquement $\rho \geqslant 0.7\rho_{\max}$.
- La puissance de calcul nécessaire est faible. On peut le programmer de manière optimisée en actualisant les résultats intermédiaires obtenus à la position t pour trouver ceux de la position $t + 1$. L'algorithme traite une courbe de 150 jours en une seconde.
- Cette technique évite de perdre les transits situés aux extrémités de la courbe : il suffit de ne travailler que sur la partie de gabarit qui reste superposée à la courbe.
- Enfin, elle se prête bien à un calcul au fils de l'eau à mesure de l'arrivée des données.

9.1.4 Modèles de transit

Les modèles de transits vont du plus simple au plus complet. Le transit élémentaire est un simple créneau carré. Bien qu'approximatif, ce modèle reste correct pour les transits de plein fouet (équatoriaux) car ils ont une signature quasiment "à fond plat". C'est suffisant quand Δf ne se démarque pas suffisamment de σ, l'écart-type de la courbe, car alors l'estimation des paramètres secondaires est impossible. Son utilisation permet de réduire à deux le nombre des paramètres car la profondeur Δf la plus vraisemblable se déduit directement de la courbe comme l'ont démontré Kovács et al. (2002) et Aigrain & Favata (2002).

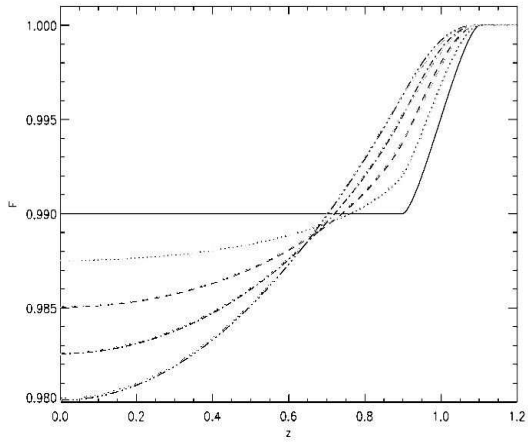


FIG. 9.2 – *Transit analytique pour un rayon planétaire $r/R_\star = 0.1$. La courbe en trait plein concerne un disque stellaire uniforme, les autres sont obtenues en incorporant différents termes de la loi de gradient de la loi d'assombrissement centre-bord. Les lignes fines à peine discernables des autres montrent l'approximation quand le flux est constant sous la surface recouverte par la planète.*

Les modèles intermédiaires sont le modèle géométrique où un petit disque planétaire occulte une partie de la surface d'un disque stellaire uniforme, et le transit en marche d'escalier utilisé par Gregory & Loredo (1992).

Le modèle le plus complet prend en compte à la fois la phase d'immersion et l'assombrissement graduel du centre vers le limbe de l'étoile. Claret (2000) a montré que la luminosité d'un point du disque d'une étoile de la séquence principale suit une loi quadratique en fonction de $\cos\theta$, l'angle entre la visée et sa surface ($\theta = 0$ au centre de l'étoile et $\theta = 90°$ au bord). Partant de là, Mandel & Agol (2002) donnent une formulation analytique du transit d'une planète devant une telle étoile, dont quelques exemples sont reproduits figure 9.2. Si la planète

est petite, on peut négliger le gradient de flux sous l'occultation. Ce modèle fournit le paramètre d'impact et le rapport des rayons quand les conditions de bruit le permettent.

9.1.5 Gabarit optimal

Le gabarit d'un mono-transit comporte le profil temporel de l'occultation entouré de deux segments constants quand la planète n'est pas devant l'étoile. Ses paramètres libres sont la longueur totale n du gabarit, et la durée d du transit inclus (voir Fig. 9.3). Nous avons conduit des essais de robustesse par injection de transits sur l'une des 1 000 courbes, qui montrent que deux durées du transit de référence, 5 h et 10 h au sein d'un gabarit de 27 h, suffisent à couvrir tous les transits entre 3h et 14h qui seraient contenus dans une courbe.

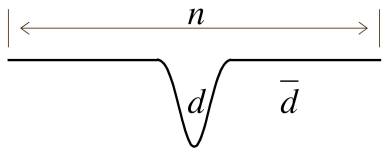


FIG. 9.3 – *Paramètres libres du gabarit.*

Il reste encore à déterminer n. Une période de calme où la planète n'est pas devant l'étoile est indispensable dans le gabarit car elle fait partie intégrante du phénomène physique étudié. Au sens de la détection, elle permet de calibrer le bruit local. Nous l'avons vérifié expérimentalement : quand n est réduit à la seule éclipse, la corrélation du transit est supérieure, mais noyée dans un bruit de détection important. De nombreuses fluctuations aléatoires de la courbe ressemblent alors à l'éclipse isolée. Si à l'inverse n est trop grand la corrélation du transit diminue car le poids de l'éclipse devient faible. D'après la table 9.1, l'optimum se situe vers $\overline{d}/n \simeq 70\%$, $\overline{d}$ étant la durée totale hors immersion.

9.1.6 Premier résultat

La détection précédente à été appliquée sur les 1 000 courbes de lumière du test en aveugle. Elle a permis d'identifier 12 événements, certains sont représentés figure 9.4. Les événements apparaissent avec un rapport signal à bruit de la détection d'environ 5σ. Sur la courbe 31, la dissymétrie alternative des événements trahit leur origine de binaire de fond à éclipses.

TAB. 9.1 – *1) Dissymétrie de la distribution des pics de détection mesurée par son 3^{ime} moment (skewness), en fonction de la longueur hors transit $\overline{d}$ du gabarit. La mesure porte sur une courbe qui contient des transits durant entre 3 h et 14 h. Un 3^{ime} moment important traduit l'émergence de pics de détection. A son maximum, le pic de détection culmine à 7σ au-dessus du continuum (cas favorable).*

$\overline{d}/n$	d=5 h	d=10 h
0.95	9	$\simeq 0$
0.90	20	2
0.80	18	10
0.60	15	8
0.40	16	7
0.20	$\simeq 0$	3

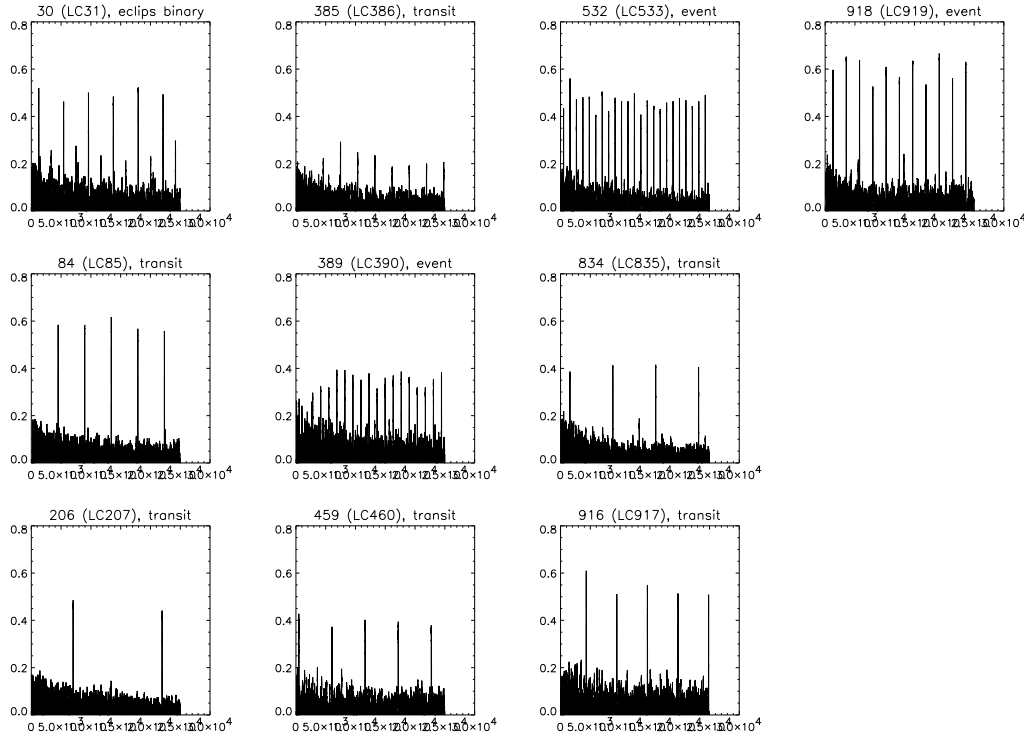


FIG. 9.4 – *Détections d'événements par simple corrélation d'une courbes de lumière avec un gabarit. L'abscisse des courbes est le numéro de point durant toute la période d'observation, qui va de 0 à 25 055. En ordonnée de chaque courbe, la valeur du coefficient de corrélation avec le gabarit.*

9.2 Le débruitage

9.2.1 Prépondérance des bruits systématiques

Les méthodes de détection de transit décrites antérieurement au test en aveugle faisaient état d'un fonctionnement courbe après courbe. Or dans les courbes de lumière du test en aveugle, le bruit est nettement dominé par des motifs collectifs aux courbes (voir Fig. 9.5). C'est donc la première des sources de bruit à traiter pour ouvrir l'accès aux informations plus fines dissimulées dans le même bruit. Nous allons donc chercher la meilleure manière d'identifier et d'exploiter les informations collectives.

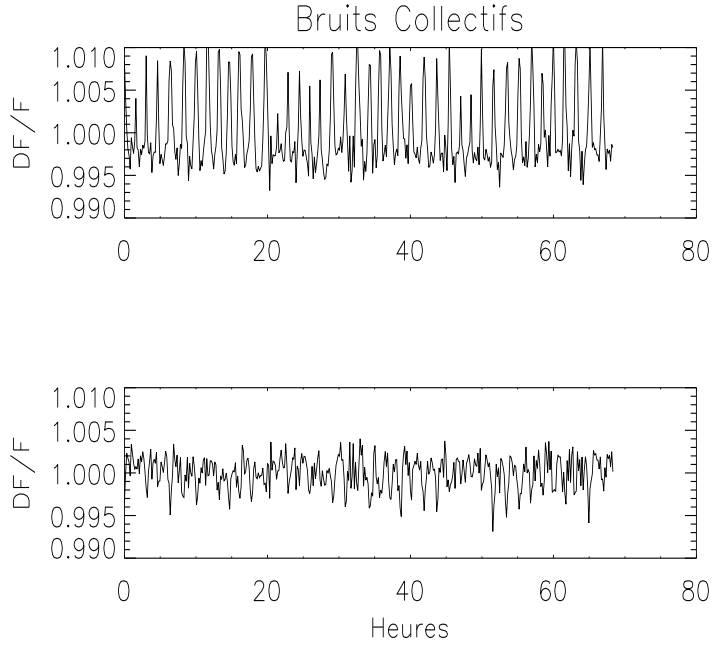


FIG. 9.5 – *Composantes systématiques. Deux courbes de lumières typiques sont représentées. Bien qu'à un degré moindre, on retrouve dans celle du bas (courbe n°5) des structures présentes dans celle du haut (courbe n°1).*

.

Nous ouvrons une parenthèse anecdotique pour illustrer par un exemple le potentiel des méthodes statistiques appliquées à un grand nombre de courbes. Il s'agit de l'extraction de la clé cryptographique d'une carte à puce par Kocher et al. (1999), à partir de courbes de consommation de courant temporelles (quelques dizaines de milliers, comme pour Corot). Les courbes étaient acquises et de manière synchrone (autre point commun avec Corot) durant le

chiffrement. Sa technique (aujourd'hui obsolète) utilisait un test d'hypothèses très judicieux portant sur 64 valeurs possibles d'une variable intermédiaire de calcul, directement liée à la valeur d'un fragment de clé. Il parvint ainsi à isoler le rôle de quelques transistors parmi les dizaines de milliers commutant pseudo-anarchiquement dans le microprocesseur.

9.2.2 Identification des modes communs

Cette tendance aux modes communs persiste quand on passe des courbes de lumière aux courbes de détection. Ces dernières ont des airs de famille, marqués par une décroissance générale du continuum de détection durant les 50 premiers jours et un pic central marqué, positif ou négatif en milieu d'observation. La figure 9.6 montre trois exemples représentatifs.

Il peut y avoir deux causes à ces ressemblances :

1. La présence d'une composante déterministe, visible ou cachée par le bruit,
2. La variation simultanée des écarts-type du bruit aléatoire.

Le pic central s'explique finalement par un phasage transitoire entre la SAA et le pic maximum de lumière diffusée dans cet exercice. Les poses les plus lumineuses sont systématiquement perdues, faisant brièvement chuter la moyenne. Ce phénomène apparaît avec presque toutes les méthodes, contraignant à ignorer les données correspondantes. On verra qu'il est naturellement pris en compte par notre processus de filtrage, et que rien dans le principe n'interdit la détection d'un transit centré à cet endroit.

Ces ressemblances sont visibles, mais suggèrent qu'il peut en exister d'autres, trop faibles pour être flagrantes mais exerçant tout de même une influence sur la détection. Nous allons les identifier toutes au moyen de statistiques portant sur la collection complète de courbes.

Un nouvel indice qui plaide en faveur de l'application préalable de la détection est que la similitude entre les ρ_i dépend peu de la variabilité stellaire qui se trouve dans les courbes de lumière. Celle-ci à donc été significativement rejetée par la détection.

La méthode élémentaire pour isoler une information constante parmi un grand nombre de courbes bruitées est d'en faire la moyenne, en espérant ainsi faire apparaître le profil commun. Considérons le modèle additif suivant :

$$\rho_i = s_i + w_i\delta + B$$

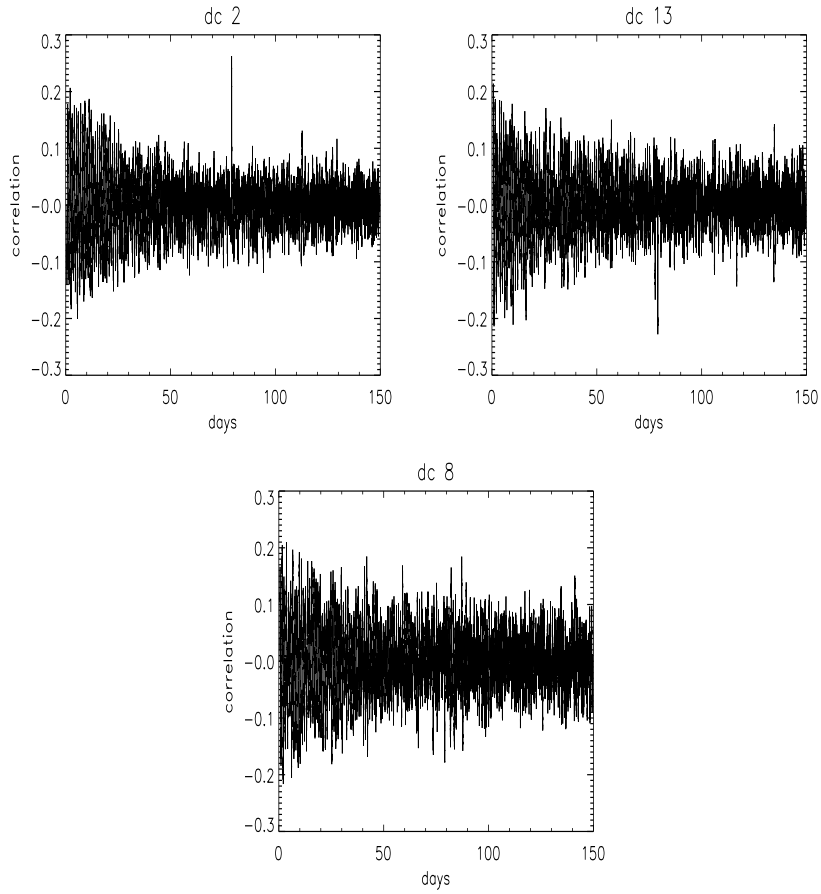


FIG. 9.6 – *3 courbes de détection présentant des motifs communs. On voit distinctement la persistance d'une tendance. Comme l'atteste le pic central cette tendance peut être positive, négative, ou à peine décelable.*

Où ρ_i est la i^{me} courbe de détection, s_i ses transits décorrélés entre les courbes, δ le motif sous-jacent pondérée affecté d'un poids w_i propre à chaque courbe. Enfin B est le bruit général de moyenne nulle. La moyenne d'ensemble sera alors égale à :

$$\langle \rho \rangle = 0 + \langle w \rangle \, \delta + 0$$

et donc proportionnelle à δ. Mais cette technique ne fonctionne pas avec les courbes du test aveugle, ce qui fait penser que w est de moyenne faible, voire nulle. Cette supposition s'est révélée exacte (après le test) car w simule un résidu de correction imparfaite de la lumière diffusée.

Il faut donc une autre méthode pour identifier δ. On a recours à une Analyse

en Composantes Principales (PCA) décrite dans l'article de Guterman et al. (2005). Le δ ainsi identifié est présenté en figure 9.7. On reconnaît le pic central et les oscillations de début d'observation. L'importance du bruit en début de détection s'explique donc par un facteur systématique plus important et non un réel bruit. En fin d'observation, l'amplitude de δ n'est pas négligeable. Un part du continuum de détection s'avère donc être déterministe. On peut donc espérer en extraire des informations.

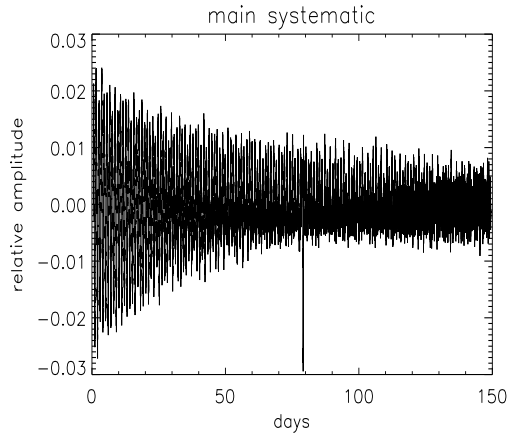


FIG. 9.7 – *Composante systématique identifiée par PCA sur 200 courbes de détection. Le vecteur à été normé. On distingue des stries à la période $T = 1.7$ h correspondant à la lumière diffusée. La méthode à donc "appris" la lumière diffusée.*

Lors du test aveugle, nous avons commencé par identifier δ à l'aide d'une technique équivalente[1]. Dans cette technique, on part de l'une des courbes choisie très typée et on la considère comme l'ébauche δ_0 de la composante commune. Cette ébauche est inévitablement biaisée. On la raffine en moyennant toutes les autres courbes préalablement normées positivement ou négativement pour égaliser leur composante suivant δ_0. Les biais sont non corrélés et s'annulent dans la moyenne, laissant l'accès à δ.

9.2.3 Obtention des courbes de vraisemblance

La technique la plus répandue pour supprimer une composante gênante est de la soustraire après l'avoir convenablement pondérée, pour ne conserver que

[1]On peut en trouver les détails sur le site de la $6^{ème}$ semaine Corot , `http ://www.ias.u-psud.fr/medoc/cw6/`

les pics de détection informatifs. Le poids w de δ dans une courbe ρ s'obtient simplement par :

$$w = \overrightarrow{\delta}.\frac{\overrightarrow{\rho}}{\|\rho\|}.$$

Mais cette opération n'est pas optimale dans notre cas car :

- Le poids w ainsi obtenu est une constante pour toute la courbe, alors que la vraie pondération devrait être de la forme $w(t)$. En effet, $\rho(t)$ dépend du bruit local et donc le même événement situé à deux instants dont les bruits ambiants sont différents peut produire deux pics différents.
- On n'a pas encore exploité l'information de dispersion des courbes de détection autour de la composante commune. On l'utilise donc maintenant pour indiquer le niveau de confiance d'un écart donné.

Dans ce but, on définit l'écart-type instantané des courbes ρ_i autour de δ :

$$\sigma^2(t) = \mathrm{v}ar(\rho(t) - w.\delta(t))$$

où w est la pondération idéale de δ dans chacune des courbes. On peut à présent convertir les pics en vraisemblance, c'est-à-dire la probabilité que le pic ne soit pas dû au bruit dans le modèle Gaussien :

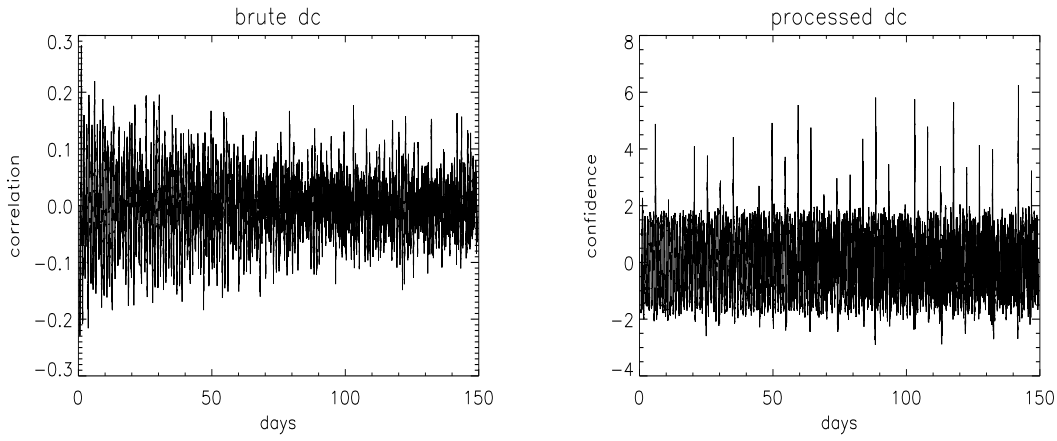
$$L_i(t) = \frac{\rho_i(t) - w_i.\delta(t)}{\sigma(t)}$$

La figure 9.8 montre un exemple de l'effet de cette conversion. La courbe de gauche est une courbe de détection brute, celle de droite est la vraisemblance correspondante. La conversion fait nettement ressortir les pics de détection. On observe l'effet régulateur de cette opération : le faux pic central de détection central à complètement disparu et l'anomalie du continuum des 50 premiers jours est éliminée, sans avoir eu besoin de les modéliser.

9.2.4 Classification des courbes de lumière

Ceci étant, l'efficacité du procédé varie significativement d'une courbe à l'autre. Le facteur identifié n'est donc pas commun à toutes. Nous avons supposé que différents bruits ont été injectés dans différentes familles de courbes pour tester les méthodes suivant différentes hypothèses. Dans certains cas, cela rend artificiellement inopérante la caractérisation collective.

Afin de se remettre dans les conditions d'application de notre méthode, nous devons regrouper les courbes par types de bruit, sans connaître ces types par



1) 2)

FIG. 9.8 – *Filtrage des modes communs. 1) courbe de corrélation contenant des pics peu visibles. 2) les pics sont traduits en indice de confiance et les transits émergent clairement. L'excès de bruit des 50 premiers jours à disparu.*

avance. Cette tâche d'identification de représentants parmi une population est précisément la fonctionnalité de l'algorithme à K-moyennes déjà décrit au §6.6 dans la réduction des masques. Les représentants sont ceux qui minimisent la distance Euclidienne de tous les groupes.

Mais cet algorithme n'est pas adapté à notre recherche d'effets systématiques. En effet, supposons deux courbes ρ_1 et $\rho_2 = -k\rho_1$, où k est un coefficient quelconque. Leur distance peut être importante bien qu'elles soient identiques pour les systématiques. On pourrait songer à utiliser comme métrique le coefficient de corrélation lui-même entre ρ_1 et ρ_2 (cf. Eq. 9.1) :

$$D = corr^2(\rho_1, \rho_2)$$

qui est insensible a k. Mais dans ce cas, c'est le déplacement des graines (cf. Fig. 6.8) au barycentre du groupe qui pose problème car il ne minimise plus cette distance.

On a recours à un algorithme rudimentaire mais efficace décrit figure 9.9. La première courbe est choisie comme 1^{er} "père", on lui attribue les "fils" qui lui ressemblent. L'une des courbes restantes est désignée comme $2^{ème}$ père et ainsi de suite. Le filtrage des composantes communes a alors lieu famille par famille.

L'ensemble de ces opérations a fait apparaître 10 détections supplémentaires en diminuant le niveau de continuum dans les courbes de détection.

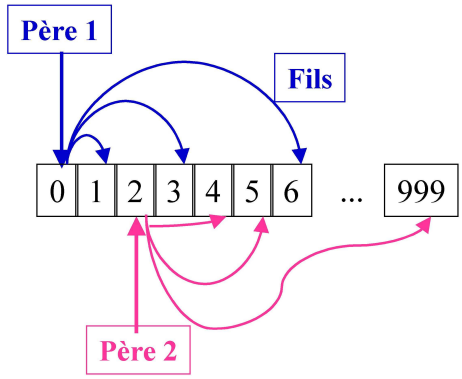


FIG. 9.9 – *Classement par famille. On choisit un premier "père". Puis ses fils sont les courbes corrélées au-delà de* $|\rho| \geqslant 0.7$*. On sélectionne alors le deuxième père parmi les courbes orphelines et ainsi de suite. On obtient un ensemble de familles dont les membres présentent des ressemblances au sens d'une composante homothétique.*

Mais dans le contexte du test en aveugle, l'effet d'apprentissage a pu involontairement éliminer une partie de la variabilité stellaire. Le résultat sera donc révisé à la baisse avec les vraies données. En effet, ce sont les mêmes profils qui sont injectés dans plusieurs courbes, après déphasage et dilatation en temps. Si ces distorsions sont trop faibles il arrive, quoique très rarement, que la variabilité se trouve brièvement en concordance dans deux courbes. Ceci est néanmoins suffisant pour que la méthode se rende compte de la ressemblance et la corrige automatiquement en tant qu'erreur systématique. Ce cas ne s'est présenté qu'une dizaine de fois, la correction n'est que locale et se limite au début de courbe, et nous n'avons volontairement pas utilisé les profils mis à jour de cette manière car cela aurait été hors de l'objectif du test. Aucun transit n'a été trouvé de cette manière, mais ce fait confirme l'efficacité de la méthode à repérer et traiter des sources de bruit inattendues.

9.2.5 Caractérisation des événements

L'étape suivante est le contrôle de la périodicité des événements détectés. On n'effectue ce contrôle que pour les principaux pics détectés. On peut alors replier les courbes suivant la période identifiée pour contrôler si leur origine est un transit ou non, dans le cas où au moins trois événements sont présents.

En superposant n transits, les barres d'erreur diminuent d'un facteur $\sqrt{n}$. Pour diminuer un peu plus le bruit, le repliement est effectué en recentrant

préalablement les événements. Le recentrage est double : en ordonnée il est nécessaire car le niveau moyen de la courbe fluctue à cause de la variabilité stellaire, en abscisse il est nécessaire car la date des événements n'est déterminée qu'avec une précision limitée par le bruit. Ce recentrage est effectué par rapport à un événement "moyen", défini comme celui qu'on aurait en l'absence de recentrage.

Afin de discriminer les transits se produisant sur l'étoile cible de l'effet des binaires à éclipses d'arrière plan, on compare 3 moyennes : celles des événements d'indice pair, impair, et total. Les éclipses dont la dissymétrie est suffisamment marquée auront des moyennes paires et impaires distinctes, réparties symétriquement autour de la moyenne générale.

La figure 9.10 montre quelques exemples de caractérisation. Les points sont les événements après recentrage, les lignes continues sont les 3 moyennes. Les barres d'erreur ne sont pas représentées. On compte une courbe à transits planétaire, une binaire à éclipses, un couple d'étoiles serrées et une planète n'ayant produit que deux transits. Malgré le "nuage" des points de mesure, la forme des transits se dessine précisément car les trois moyennes sont très proches.

9.3 Résultats

Par rapport aux résultats obtenus après la seule phase de détection (12 événements), le processus de débruitage collectif conduit à 10 événements supplémentaires (+80%). L'amélioration est donc tangible pour le détecteur par corrélation.

Pour vérifier que l'amélioration ne se limite pas aux événement "faciles", il faut comparer notre résultat avec celui des autres équipes. Ce bilan est explicité dans l'article de Moutou et al. (2005) et synthétisé dans la table 9.2 ; notre résultat est médian. Les transits concernés sont bien d'un niveau égal à ceux des autres méthodes ce qui, compte tenu de la performance modeste du détecteur employé, souligne bien l'intérêt du débruitage par identification des déterminismes collectifs cachés.

Les transits vrais détectés à l'occasion du test sont les mêmes quelle que soit la méthode utilisée. Ceci donne une idée de la sensibilité de détection de Corot . A l'inverse, les fausses détections dues au bruit sont différentes entre les méthodes, donc la comparaison semble être un critère discriminatoire. On note aussi que la variabilité stellaire ne semble pas gênante pour la détection si elle reste inférieure à 0.5%. La détectivité de Corot devrait rendre possible la

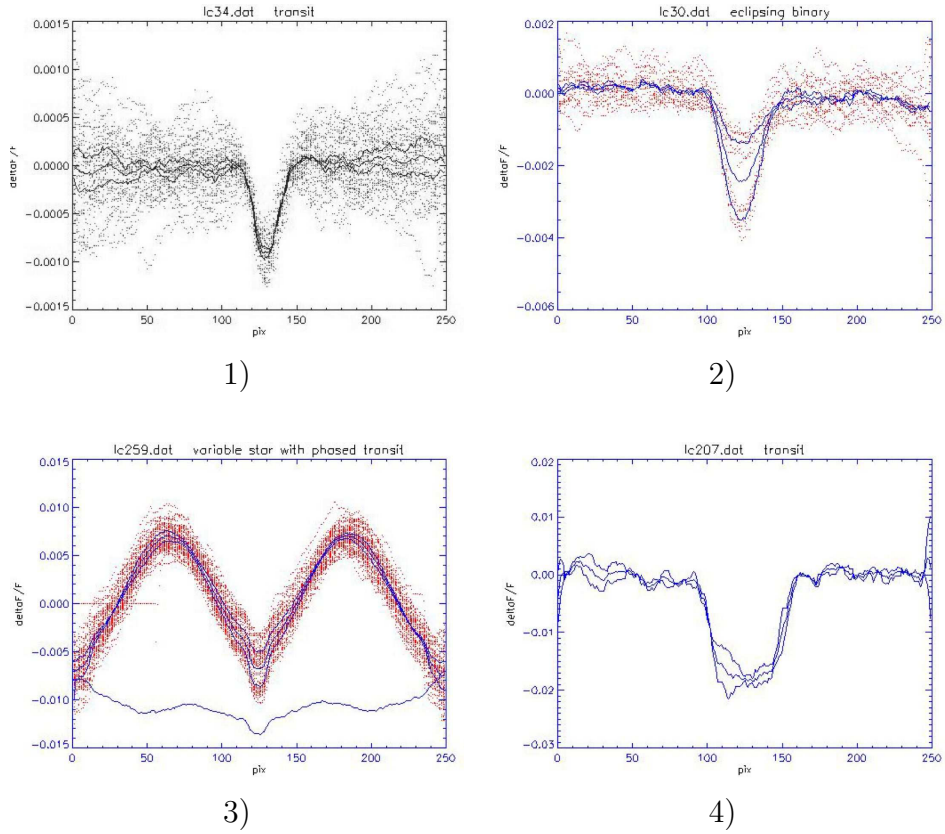


FIG. 9.10 – *Discrimination des événements. Les points sont ceux des événement détectés, les courbes sont les moyennes des événements d'indice pair, impair et la moyenne générale. 1) Courbe avec transits planétaire. La forme du transit apparaît distinctement, les trois moyennes sont proches. 2) Une étoile de fond binaire à éclipses 3) Une étoile double en orbite serrée. Il s'agit d'un objet double en rotation, et non plus d'éclipses, 4) Une planète n'ayant transité que deux fois.*

TAB. 9.2 – *Résultats comparatifs. Notre méthode (équipe 1) se situe dans la moyenne. Elle évite les fausses détections grâce à une phase d'analyse visuelle semi-manuelle. Les détecteurs arrivant en tête utilisent le repliement.*

Equipe	1	2	3	4	5	Parmi
Transits	12	11	15	12	14	20
Autres événements	10	6	10	12	12	16
Fausses détections	0	1	0	3	5	

découverte de planètes de $1.1R_{\oplus}$, autour d'étoiles naines $M0$ pourvu que leur période soit inférieure à 3 jours.

Il est possible que l'apport de notre approche soit moindre avec les vraies données de Corot qu'avec les données simulées du test en aveugle. En effet, les bruits expérimentaux seront plus efficacement traités par l'EXOPIPE à l'aide de modèles spécifiques, et peut-être évolutifs. Ce faisant, il accomplira en amont une partie du travail de notre méthode car ces bruits contiennent une grande part des déterminismes cachés. Mais elle restera utile pour traquer les bruits résiduels dus à l'imperfection des modèles.

En revanche notre méthode peut faire face à des situations imprévues, comme on l'a vu dans l'exemple du pic central de fausse détection (cf. Fig. 9.6). Cet artefact d'envergure majeure situé en milieu d'observation est causé par la conjonction entre la SAA et le pic de lumière diurne. Il gêne la majorité des équipes en lice, qui n'ont d'autre recours que d'ignorer cette fausse détection à l'aide d'un traitement ad hoc, disqualifiant ainsi les données correspondantes. Au contraire avec notre méthode, le pic s'évanouit de lui-même sans faire l'objet d'aucun traitement particulier, permettant à l'information occultée de revenir d'elle-même au premier plan.

De la même façon, la mise à jour involontaire de quelques profils de variabilité stellaire contenus dans les courbes du test en aveugle montre la capacité de notre approche à réagir rationnellement aux sources de bruit inattendues.

Le pic est commun à la plupart des méthodes car elles réagissent toutes aux mêmes facteurs. Elles sont certainement aussi gênées par les autres effets systématiques moins visibles que nous avons mis à jour. Quelle que soit la méthode, la prospection et la suppression de ces déterminismes qui sinon restent considérés comme du bruit de détection ferait certainement progresser la détectivité de Corot .

La contribution du débruitage collectif à été confirmée depuis par Tamuz et al. (2005) qui utilisent une stratégie similaire, adaptée aux conditions de l'expérience OGLE. Ayant identifié une source de nuisance, ils recherchent dans toutes les courbes sa meilleure pondération en supposant qu'elle produit un effet linéaire ou quadratique. Leur calcul tient compte des incertitudes individuelles de mesure. Par exemple pour l'extinction atmosphérique, supposons que $a(t)$ soit l'épaisseur d'air traversée à l'instant t. Si $s_i(t)$ est la $i^{ème}$ courbe de lumière (t est un temps discret), on cherche la pondération c_i qui y accorde le mieux $a(t)$. Il s'agit de minimiser la fonction :

$$S_i^2 = \sum_t \frac{\left(s_i(t) - c_i a(t)\right)^2}{\sigma_i^2(t)}$$

où $\sigma_i(t)$ est l'erreur de mesure à l'instant t. Cette équation admet une solution

formelle. Une fois en possession des pondérations c_i de toutes les courbes, on peut raffiner rétroactivement le modèle de perturbation $a(t)$ en $a'(t)$ car son effet n'est pas forcément linéaire, notamment à cause des changements météorologiques. On obtient :

$$a'(t) = \frac{\sum_i s_i(t)c_i/\sigma_i^2(t)}{\sum_i c_i/\sigma_i^2(t)}$$

Tamuz et al. (2005) généralisent leur technique à plusieurs effets indépendant linéaires ou quadratiques, notamment les phases de la lune et la position de l'étoile sur le CCD.

Leur méthode s'avère plus efficace que l'analyse en composantes principales, mais nécessite d'avoir identifié les sources de bruit et de disposer des erreurs de mesure. Ces spécificités sont complémentaires à la nôtre et suggèrent l'avantage qu'aurait un usage conjoint. A la réception des données Corot , on conçoit qu'il soit raisonnable de commencer par faire un "état des lieux" du bruit par exemple avec l'approche que nous proposons, puis d'agir sur les bruits que l'on sait modéliser par des méthodes plus ciblées.

9.4 Perspectives

Du côté détection, on peut combiner une approche collective non plus avec notre détecteur basique, mais avec les détecteurs fonctionnant par repliement des courbes de lumière qui donnent actuellement les meilleurs résultats. On pourrait de plus exploiter l'information de la symétrie intrinsèque des transits à l'aide d'un auto-repliement supplémentaire par rapport à leur centre.

Un premier perfectionnement du débruitage collectif est son extension à des sources de bruit multiples et locales. En effet, celle-ci est compatible avec l'identification de plusieurs caractéristiques communes même si elles n'ont pas les mêmes poids relatifs dans les courbes. Elle permettrait seule, ou secondée d'un classement par familles, de déduire la combinaison de certaines perturbations, sans connaissance préalable des courbes concernées.

L'application d'une approche collective à la variabilité stellaire peut s'avérer d'un grand intérêt. Sa cible ne serait plus la concordance temporelle des signaux, mais l'identification et la caractérisation statistique des familles de variabilité. On pourrait par exemple dégager une matrice de covariance représentative de chaque famille pour exploiter pleinement les capacités du filtrage adapté.

9.5 Troisième publication Guterman et al. (2005), SPIE.

Improving transit detection with collective light curves information

Guterman P.[a], Barge P.[a], Llebaria A.[a], Quentin C .[a]

[a]Laboratoire Astrophysique de Marseille(CNRS), 13776 BP8 Marseille Cedex12, France

ABSTRACT

The search for planetary transits in star light-curves can be improved in an non standard way applying appropriate filtering of the systematic effects just after the detection step. The procedure has been tested using a set of light curves simulated in the context of the CoRoT space mission. The level of the continuum in the detection curves is significantly lowered when compared to other standard approaches, a property we use to reduce false alarm. Ambiguities may originate in unexpected effects that combine instrumental and environmental factors. In a large set of synchronous light curves collective behaviours permit to identify systematic effects against which the detected events are compared. We estimate a significance of our detections and show that with our procedure the number of true detections is increased by more than 80% (22 events detected over the 36 injected ones). In spite of its simplicity, our method scores quite well (average results) when compared to the other methods used for the CoRoT "blind test" exercice by Moutou et al.[1]

Keywords: corot, transit detection, survey, pattern analysis , stellar photometry

1. INTRODUCTION

1.1. Corot

High precision stellar photometry permits to detect extra-solar planets by the transit method. Indeed, the transit of a planet in front of the disk of its parent star results in a photometric signal (a weak, short and periodic decrease of the received photon flux F) that can be detected in stellar light-curves. The relative amplitude of the signal ranges from 10^{-2} for giant planets like Jupiter to 10^{-4} for a terrestrial planet; its periodicity P and duration Δt are related to the orbital motion of the planets ($P \simeq$ weeks to months or years, and $\Delta t \simeq 1-10hrs$). The CoRoT small satellite project is a space mission which is half devoted to search for telluric extra-solar planets (the other half is devoted to stellar seismology). A description[2] can be found at link *http://smsc.cnes.fr/COROT/* It will permits the detection of photon flux variations about 7.10^{-4} in a one our integration time, compatible with planetary transits on stars whose magnitude ranges from 12 to 15.5 . CoRoT should be able to detect 10-40 terrestrial planets and several of "hot Jupiters" dependent on the a priori hypotheses on the existence of the planetary systems. Space missions such as CoRoT will provide very large number of light-curves, which will require optimized and automated detection algorithms for the processing of the data. Various detection algorithms have been developed so far, but an important question is also to get free at best from instrumental noises and systematics. This is the point we want to address in the present paper, using the collective information from the large number of available light-curves in a field of view. A new approach of the problem in which detection precedes the filtering procedure.

1.2. Signal and Noises

The signal is a nearly constant flux F marked by the slight dips due to the possible transit of a planet in front of the star disk. The typical value of the flux for a star with mv=14 is 2.10^6 phe$^-$ in a 512s exposure time. The signal is affected by various noises and perturbers namely, by decreasing order of importance:

- The **photon noise**. It is proportional to the square root of the number n of photons received from the target star. Of course the larger n is, the better s/n is. However, due to PSF overlapping in crowded field, photons of the target are mixed with non-informative photons of the background stars so that s/n is decreased.

- The **scattered light**. Rejection of light reflected from the Earth by a high performance baffle allows to reduce the stray-light by a factor of 10^{-13}. The residual disturbance is strongly modulated at the orbital period of the satellite. The mean induced variations can be corrected but not its associated photon noise. In the simulated data used in our tests the mean level correction was imperfectly modeled.

- The intrinsic **variability** of stars. Up to now, only the variability of the Sun is well known, mainly through the observations of a complete activity cycle with the SoHO satellite. On the other hand the variability of stars others than the Sun remains poorly known (it will be explored with CoRoT). It has been mimiced in our simulated data using the Virgo-SoHO data with appropriate scaling and extrapolations as described by Moutou et al.[1]

- **Jitter and breathing**. Jitter noise is the residue of fast pointing corrections which causes random variations of the signal (with CoRoT a typical 0.1 pixel amplitude is expected). Breathing is due to the thermo-elastic deformations of the telescope structure at the orbital period (much longer than jitter period) and can be approximated by a simple defocus. The residuals after correction of these perturbations are not simulated in the present set of our test light curves.

- The **readout** noise of the CCD. It directly contribute to the total variance, like dark current. The non uniformity of pixel or sub-pixel response acts in combination with jitter.

- **Cosmic rays.** The glitches (local saturations) they form on the CCD, may imply a consequent loss of information for the target stars. Saturated exposures are rejected on-board and produce gaps in the data localized at SAA crossing.

1.3. The various methods used on the simulated light-curves

Various algorithms have been developed to detect planetary transits in star light-curves. In the bayesian approach of Defaÿ et al.,[3] the most probable period of the transit is estimated thanks to maximum likelihood and the transit signal is reconstructed using the Fourier coefficients of the fundamental harmonics. The well known matched filter can be used with a large collection of reference transits as those computed by Jenkins[4] in the general case of a planet orbiting a binary star. Finally combining unknown periodicity and a priori shape Aigrain et al.[5] and Kovács et al.[6] fold the signal to improve S/N and make detection with a box fitting algorithm using χ^2-test. These authors also deduce the optimal in-transit and out-transit level from the data-set itself. No detection algorithms based on Fourier Transform have been used because the low energy of transits is spread down into many harmonics because of it's temporal briefness.

The performances of these various methods were compared on simulated COROT light curves by Moutou et al.[1] They require a preliminary denoising step to avoid spurious detections due to straylight or star variability. The main tools used at this stage were: simple thresholding, median filter, non linear filter based on a structuring element, polynomial local fitting for removal of the stellar variability, subtraction of long term variations or fitting with a family of sinusoids. On the other hand, in the context of the Kepler mission, Jenkins[7] removes stellar variability using wavelets. In each case, gaps in the data were processed differently, being interpolated or not.

Two of the 1000 light curves of the blind exercise are plotted in Fig. 1. Only the right curve (no 34) contains transits. But they are concealed as a set of spikes embedded in noise. Nevertheless they have been found by all the different concurrent methods (see Moutou et al.[1]).

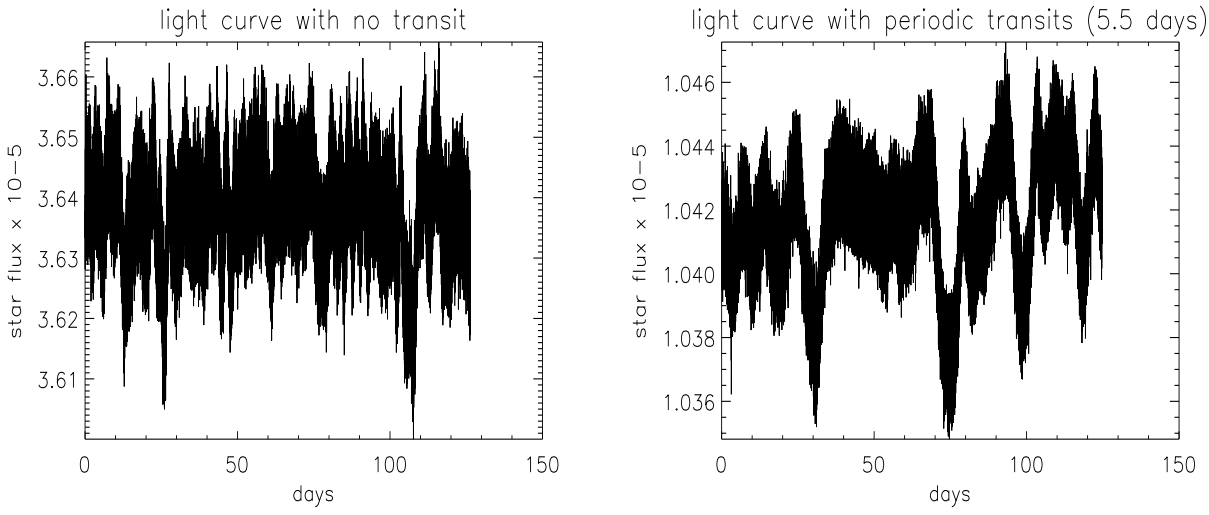


Figure 1. Raw light curve number 1 (left) and 34(right). Only $lc34$ contains transits. See main text for details.

2. THE PROPOSED METHOD

Our method referred as team 1 in the synthesis by Moutou et al.[1] starts by the detection step before any denoising. This is in contrast with all the other methods presented in referred paper.[1] They pre-process the curves to reduce the noise level in detection. In our method, detection is performed directly on the raw light curves (lc) transforming them into temporal detection curves (dc) of same lenght. Then the dc s are analyzed to bring out the unique features of each transit discarding the common noise caracteristics. The common signature of the systematics can be spotted thanks to the common behaviors across the whole set of dc' s. After discriminating analysis we obtain the confidence score for each data point in the detection space against the whole set. Finally, the list of transit candidates is sorted out and the periodicity criterion is used to decide the more likely ones.

First we will spend a while defining the two steps, namely *denoising* and *detection*.

De-noising consists in applying algorithms for reducing the noise level, in time or in frequency space. The result of the denoising step is a new light-curve lc', which is smoothed with respect to the initial lc. Some noise components are well defined like Poisson noise but others like stellar variability remain hypothetical but can be checked afterwards. Therefore filtering procedures can never be perfectly adapted.

Detecting consists in searching for periodic transit like signals in the light-curves. The detection algorithms converts lc' s into dc' s. Each point of a dc indicates the chance that the lc contains a transit at corresponding time. Our detector is correlation between the light curve and a reference transit.

The most accurate information must be used first, that is our strategy. Because transits shapes are better known than noise effects, detection must arrive first. Mandel & Agol[8] had elaborated a precise analytic expression. The main parameters are depht and duration. Also the luminosity gradient from center to limb of the star surface is considered. One example of compliance with real stars is shown Fig.2.

As said before, the lc noises are known with less precision, in particular the star activity whose models are influenced by the Sun. An initial filtering of light curves will affect all possible included transits, resulting in a degradation for later detections. At the opposite, a leading detection will be performed using a collection of transits. The closest set to the true transit corresponds to the highest signal to noise ratio in detection. Our trial

show that 2 references of transit are sufficient to cover the full duration range. One of 35 pixels (5 hours) and one of 70 pixels(10 hours). The robustness of our correlation based detection criteria avoids any pre-process prior to detection.

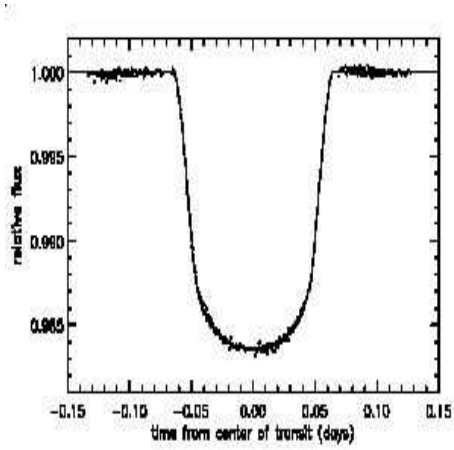


Figure 2. A planet occulting star $hd_2 09458$ observed by the Hubble Space Telescope. Experimental points are perfectly sticking to the model over plotted.

Common methods for light-curve de-noising are based on time averaging, frequency filtering, interpolation/removal of data or morphological non linear filtering. These methods focus on reducing noise levels but must also avoid to perturb the transit signals. Perturbations are very weak when the noise frequencies are far from the frequencies associated to the transits. However, as the goal is to reduce at best the noise contributions, a natural tendency is to get closer to the noise/transit limit. For instance, in the case of CoRoT we are interested in removing harmonics at the orbital period of the satellite (100 minutes) which modulate the stray-ligth and the PSF variations. Although these periods are quite far from standard transit durations (several hours), pollutions of the transit signals are still possible: (i) via high frequencies, if working in Fourier space; (ii) via small changes in the relative flux values due to the correction of the stray-light modulations.

The inconvenience of a pre-processing of the data (prior to detection) can be illustrated in the case of CoRoT. The stray-light from the Earth is an important source of noise modulated by the orbital motion of the satellite, but can be removed after simple orbital averaging. In the case of a strong transit, detection can be different following the light-curves are corrected or not from the stray-light variations. The results we get in the case of CoRoT are plotted as correlation curves in Fig. 3. The left dc is the result of detection applied to the raw lc number 30. Right dc corresponds to the same lc, but smoothed at orbital period. Detection seems less effective if the stray-light modulations are removed from the light-curves than if they are not. In the raw case, the detection accuracy is sufficient to notice the 2 different eclipses of a background binary star. In the second case, the smoothed lc seems nicer at naked eye, but the dc reveals a loss in signal to noise despite higher correlation pics. This is likely due to the fact that in preprocessed light-curves the signal is "eroded" and patterns fitting to the transits reference become more numerous.

2.1. Detection

Our method is a simple statistical correlation between the data and a reference signal. We use the appropriate correlation coefficient:

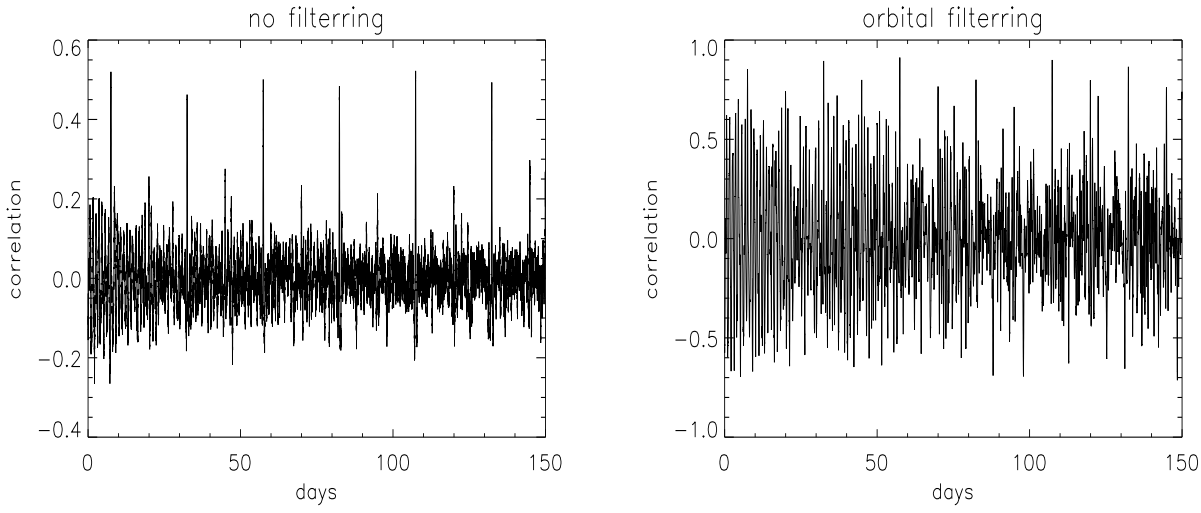


Figure 3. Detection of transit in a light curve without any pre-processing (left) and after removing the stray-light modulations (right). Detection pics are higher for pre-processed light-curves but S/N is higher for unprocessed light-curves.

$$\rho = \frac{\overline{xy} - \overline{x}.\overline{y}}{\sqrt{\left(\overline{x^2} - \overline{x}^2\right)\left(\overline{y^2} - \overline{y}^2\right)}}$$

where x is the measured series, y the model and $\overline{x}, \overline{y}$ their mean values. In our problem, x is any segment of the light curve and y a reference transit of same width. In the (x, y) plane we get a dot cloud which is flat when ρ is large, what means that lc and the reference are in good accordance at the time of x. ρ is insensitive to scales and offsets which just inflates or tilts the dots cloud, a useful property which avoids normalization between the different light-curves. The number of the unknown parameters is reduced to three: period, phase and duration.
Another advantage of the above correlation is that it can manage data gaps without interpolating missing points. This avoids artifacts and limits possible false alarms. The values of ρ range in $[-1, 1]$, whatever the flux or the amplitude of lc. This will help for further linear processing. It also copes with problem of border points exclusion ($t < l/2$ and $t > l/2$) in the time extent of the transit of reference. Light-curves with data gaps will have lower ρ values.

Detection curves dc are constructed by plotting, for each light-curve lc, the point $dc(t) = \rho(lc\,[t - l/2, t + l/2]\,, g)$ as a function of time; g denotes the reference transit (template) of length $l + 1$. This deals with the phase of the event, reducing the number of remaining free parameters at two: period and duration. As transit model we use the quadratic expression of Mandel & Agol[8] with recommended coefficients, (γ_1 and γ_2) equal to 0.5.

In fact, a transit of reference has an additional free parameter that cannot be assumed to be zero: the total time extent of the transit model $\bar{l}$ which include the transit duration l plus two out of eclipse durations, just before and just after the eclipse itself. To assess the more appropriate $\bar{l}$ we led several tries which are summarized in table 1. Paradoxically, we find that a good transit model must contain out of eclipse durations up to four time longer than the transit duration itself. This is because a full flux just before and just after the transit itself are an integral part of the planetary occultation phenomenon. Such a model brings more complete information on the transit events than the standard single dip model. Zooming Fig. 3, we found, indeed, that

negative pics with smaller amplitude are bordering detection pics; this signature corresponds to the associated anti-correlations between the transit model and the transit event when in temporal quadrature.

Table 1. Performance of detection as a function of the reference transit model (duration of the "prelude" and "coda" of the transit). We measure the skewness, with highest values for a noise histogram with a secondary bump. Both 35 pixel (5h) and 70 pixel (10h) eclipses peak around 70% of required non-transit time. We kept this value. The detection signal to noise ratio reaches 8.1 (Resp. 6.5) in mentioned favorable case.

l	95	90	80	60	40	20
$tr = 35$	9	20	18	15	16	$\simeq 0$
$tr = 70$	$\simeq 0$	2	10	8	7	3

2.2. Filtering Detection Curves

We applied the described detection for the whole set of 1000 curves. Except for a few tens, the detection curves keep noisy. But, comparing them in Fig. 4 shows obviously dominating common characteristics. We can reasonably assume that the common aspect of all these curves is originated by 1) a deterministic common embedded pattern and 2) a common distribution of random noise. Both items are discussed in this paper. An encouraging sign is that the similitude between dc' s do not depend much on star activity level. Using the synchronized acquisitions, we will identify and treat those points, each curve learning from all the others.

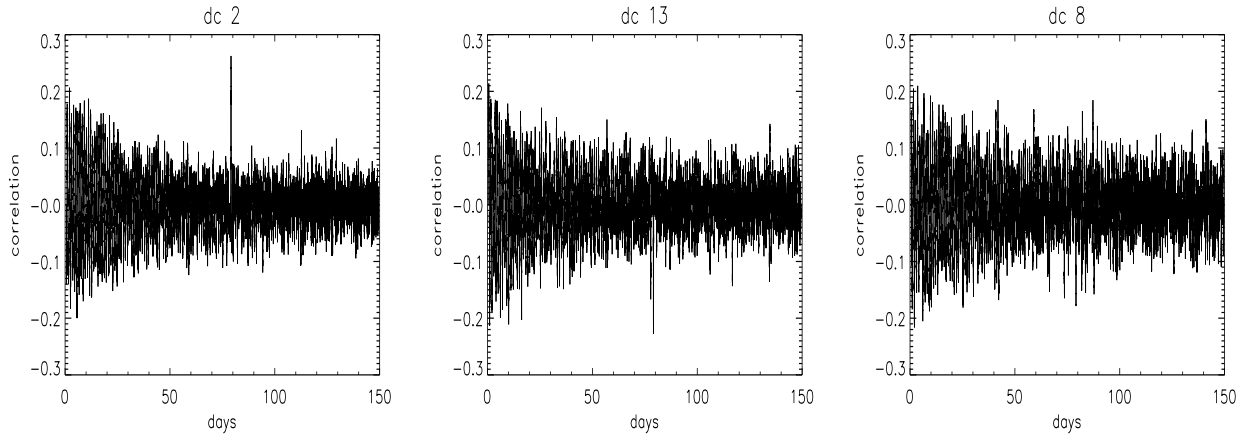


Figure 4. Systematics and trends in standard deviation. 3 representative detection curves are represented. Their similar aspect can have 4 possible sources 1) The dominating visible artifacts. They are deterministic and even sometimes known, for example the deep spike at curves center is due to local synchronism between SAA cancellation and daytime higher level. This pic is either positive or negative revealing an algebraic weight. 2) The common behavior of error bars, as marked at curves beginning. 3) Suspected tiny pattern lying under noise level, like possibly in curves second half. 4) real random noise.

To cancel uncorrelated noises and conserve the searched constant, one would naturally average dc' s all together. But it doesn't work. We suspected that, despite a constant looking at macro-scale, the added noise is actually homothetic with null average when looking closer. After the test, this was confirmed as due to simulating unperfected corrections. We use the next model to address all possible common mode variations:

$$dc_i = s_i + k_i.\delta + n$$

dc_i is the ith detection curve. All elements are unknown but have different properties: s is the searched detection signal made of pics, δ the systematic, k_i its individual weight and n the centered random noise. These properties are sufficient to derive the missing information with sufficient accuracy. Our model effectively explains that $\langle dc \rangle = 0$.

A sound tool to derive at the same time an unknown pattern and all its weights is the principal component analysis (PCA), described by Press et al.[9] In our case, we search for a constant unitary vector $\overrightarrow{\delta}$, present in all $\overrightarrow{dc}$, which accounts for most part of the variations, simply by appropriate weights. In other words the projections of $\overrightarrow{dc}$' s over $\overrightarrow{\delta}$ show a maximal variance. We can note this:

$$\overrightarrow{\delta} \,/\, \max E(\overrightarrow{dc}.\overrightarrow{\delta})^2 = \max \overrightarrow{\delta}^{\mathrm{t}}.G.\overrightarrow{\delta} \tag{1}$$

with $G = E(\overrightarrow{dc}.\overrightarrow{dc}^{\mathrm{t}})$ the covariance matrix of detection curves. Writing this equation in the eigen base it turns that we search $\{x_i\}$ the coordinates of $\overrightarrow{\delta}$ in eigen base such that :

$$\begin{cases} \{x_1, \dots x_i\} & \text{maximizes} \quad \sum \lambda_i x_i^2 \\ \text{and} & \\ \| \overrightarrow{\delta} \|^2 = 1 & \end{cases} \tag{2}$$

the λ_i are the eigen values. The eigen base is sorted by decreasing order of λ_i. The solution is $x_1 = 1, x_2, \dots = 0$. Therefore, $\overrightarrow{\delta}$ appears to be the first eigen vector of G. Weights k_i are deduced by scalar products $k_i = \overrightarrow{dc_i}.\overrightarrow{\delta}$. In practice, all dc' s are as a preventive centered and normed to avoir round-off instability in matrix diagonalisation. The result is shown in Fig.5. The found pattern looks clearly compliant with the detection curves of Fig.4.

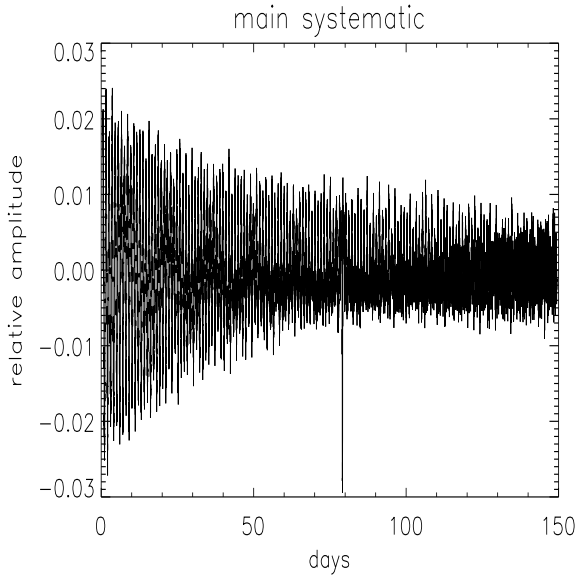


Figure 5. This curve results from the PCA between 200 detection curves (see the main text). The vector has been normed.

The central spike is due to SAA and daytime synchronism. Its depth is of same order than the amplitude at the begin of the run. So we can expect to remove a significant part of initial continuum in dc' s. The right part

of the pattern is not negligible, so we can again expect some improvement in dc' s end continuum, despite it seems only made of random continuum.

We can now remove $k_i \overrightarrow{\delta}$ from the detection signal to recover only the signal pics. But doing a simple subtraction of $k_i.\delta$ from dc_i is worthless because 1) two pics of same dc can nevertheless show different distances to $k_i \overrightarrow{\delta}$. This results from their different weighting due to their two different level of noise in original lc series at their 2 different epochs. And 2) we still ignore what confidence can be granted to a particular deviation from one pixel $dc_i(t)$ to all pixels $dc_{j\neq i}(t)$. To derive a confidence level, we first need the instantly spreading of curves around δ. At time t:

$$\sigma^2(t) = \mathrm{var}\left(dc(t) - k.\delta(t)\right)$$

We then convert pics into their significance, that's to say the probability that a pic is not due to chance in the Gaussian model:

$$L(t) = \frac{dc(t) - k.\delta(t)}{\sigma(t)}$$

Figure 6 shows an example of the regulating effect produced on noise continuum and shows how pics can emerge from noise. In the blind test we also used an equivalent technique[10] with more details at link *http://www.ias.u-psud.fr/medoc/cw6/*. We first located the common pattern by working on short sub-series of lc' s. We start by a first guess δ, then refine it by cancellation of uncorrelated biases. Then we move to PCA.

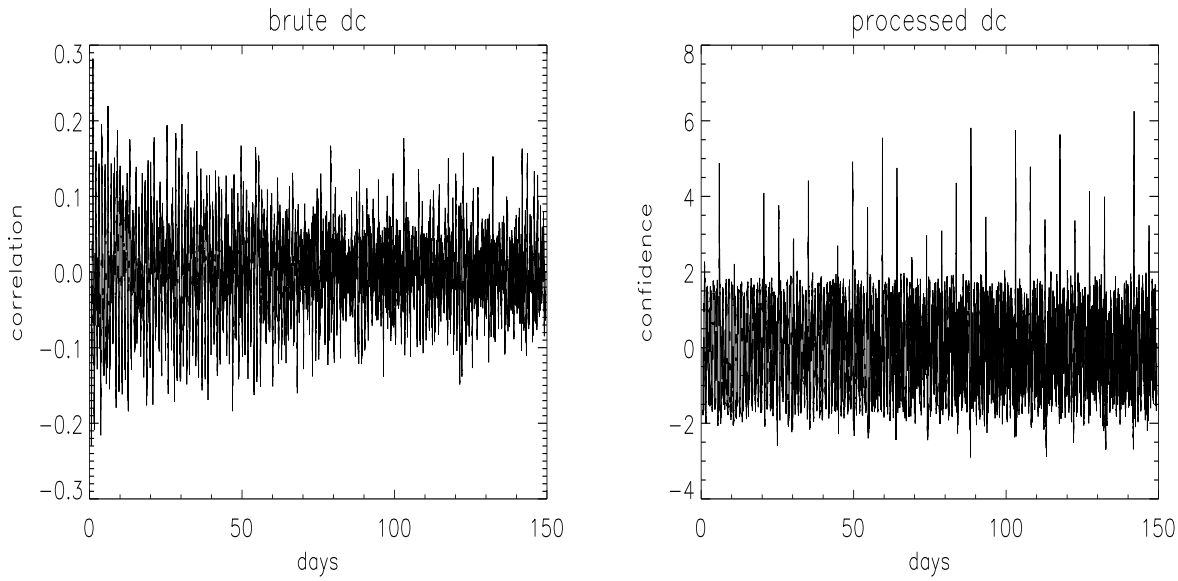


Figure 6. PCA effect. Left a raw detection curve. Abscissa are the days, ordinates is the correlation. Right the same one converted into confidence level of pics. This example clearly shows a regulation and noise lowering effect, allowing events to rise out from continuum.

Actually the efficiency is different for each curves, suggesting distinct embedded instrumental noise scheme for sake of comparison. But distinct patterns are no more systematics and will not be present in real CoRoT data. So we replayed the de-noising after a basic classification. This classification choses an arbitrary 'father', then his sons are the most correlated lc. We iterate with next father and so on. This improved the result, but the

learning effect unwillingly removed part of star activity at the same time. This comes from scarcity of star activity measurements which obliged to put close patterns in several curves.

This way we identify 12 more periodical events. At the end, all dc' s are sorted out by order of interest and the most interesting examined one by one. We control the strict periodicity of transit candidates, possibly guided by the Fourier transform, and fold lc accordingly to distinguish from non-transit events. Main confusion sources are grazing background binaries which behave like a monolithic rotating object, or background eclipsing binaries which alternate 2 different true transits.

2.3. Results

We detect 12 events after the detection step. The proposed de-noising rises this number up to 22 overall events. The repartition is 12 transits found among 22 and 10 other events over 16. The physical origin of 5 of the other events was wrong, but this identification was not part of our study. The result for all teams is resumed in table. 2 and detailed by Moutou et al.[1] The detected and not detected transits are common to all method, hence giving an idea of detection sensibility.

Table 2. Comparative results. Our method rank on the average. A given false positive is never detected twice, which gives advantage to the complementarity of the methods.

Team	1	2	3	4	5	Total
Tansits	12	11	15	12	14	20
Other events	10	6	10	12	12	16
False positives	0	1	0	3	5	

Our result ranks on the average. This is interesting, in term of improvement regarding the simplicity of the detection criteria. The interrest of such a collective approach has been confirmed by Tamuz et al.[11] which present another technique for the OGLE ground based survey (they correct linear systematic perturbations in light curves like atmosphere transparence. Their technique gives better results that the PCA, thanks to the use of measurements uncertainties).

3. CONCLUSION AND PERSPECTIVES

We have demonstrated the interest of performing detection before filtering. By first exploiting the most accurate information (the transit shape), the detection improves better the signal to noise ratio. Hence the risk of artifact caused by a filtering which transforms the signal is reduced. The detection also deals with the missing data without interpolating it. We additionally show that properly normed detection curves can be processed by linear operations, as a direct signal. This gives benefit of identifying disturbing systematic effects by the study of collective dc' s behaviors, without need of a model. This is an advantage when systematic effects can not be known and modeled with a sufficient accuracy. This approach is in fact complementary with the more classical techniques, being able to exploit the global information from the big amount of light curves obtained by COROT .

Three improvements are under study 1) The early folding of curves should improve the detection signal to noise by a factor up to $\sqrt{n}$ for an n-times folding. This is challenging leading roughly to a 3000 times computation overhead (today 1 sec/curve). Nevertheless, avoiding some redundancies in computation could reduce it to 300. For instance we could keep intermediate results in sliding correlations, compute all harmonics at same time, or even accumulate the daily data as they arrive slowly. 2) The extension of PCA to several independent cumulated patterns. 3) The use of a more specialized detector.

REFERENCES

1. C. Moutou, F. Pont, P. Barge, S. Aigrain, M. Auvergne, D. Blouin, R. Cautain, A. Erikson, V. Guis, P. Guterman, M. Irwin, A. F. Lanza, D. Queloz, H. Rauer, H. Voss, and S. Zucker, "Comparative blind test of five planetary transit detection algorithms on realistic synthetic light curves." accepted in *Astronomy and Astrophysics*, 2005.
2. "Corot: Cnes web site." http://smsc.cnes.fr/COROT/.
3. C. Defaÿ, M. Deleuil, and P. Barge, "A bayesian method for the detection of planetary transits," *Astronomy and Astrophysics* **365**, pp. 330–340, Jan. 2001.
4. J. M. Jenkins, L. R. Doyle, and D. K. Cullers, "A matched filter method for ground-based sub-noise detection of terrestrial extrasolar planets in eclipsing binaries: Application to cm draconis," *Icarus* **119**, pp. 244–260, Feb. 1996.
5. S. Aigrain and M. Irwin, "Practical planet prospecting," *Monthly Notices of the Royal Astronomical Society* **350**, pp. 331–345, May 2004.
6. G. Kovács, S. Zucker, and T. Mazeh, "A box-fitting algorithm in the search for periodic transits," *Astronomy and Astrophysics* **391**, pp. 369–377, Aug. 2002.
7. J. M. Jenkins, "The impact of solar-like variability on the detectability of transiting terrestrial planets," *The Astrophysical Journal* **575**, pp. 493–505, Aug. 2002.
8. K. Mandel and E. Agol, "Analytic light curves for planetary transit searches," *The Astrophysical Journal* **580**, pp. L171–L175, Dec. 2002.
9. W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical recipes in C - The art of scientific computation*, ISBN 0-521-43108-5, Cambridge University Press, 2nd ed., 1992.
10. "Corot week 6." http://www.ias.u-psud.fr/medoc/cw6/index.php?action=presentlist.
11. O. Tamuz, T. Mazeh, and S. Zucker, "Correcting systematic effects in a large set of photometric light curves," *Monthly Notices of the Royal Astronomical Society* **356**, pp. 1466–1470, Feb. 2005.
12. B. Tingley, "A rigorous comparison of different planet detection algorithms," *Astronomy and Astrophysics* **403**, pp. 329–337, May 2003.
13. B. Tingley, "Improvements to existing transit detection algorithms and their comparison," *Astronomy and Astrophysics* **408**, pp. L5–L7, Sept. 2003.

Chapitre 10

Conclusion

La mission Corot mettra bientôt en oeuvre depuis l'espace la méthode des occultations pour détecter des planètes extrasolaires. Corot mesurera en continu le flux de 60 000 étoiles et permettra d'atteindre la précision nécessaire pour détecter des planètes de taille tellurique. Corot devrait ainsi lever un peu plus le voile sur les mécanismes de formation planétaire et ouvrir la voie vers l'exobiologie en repérant, peut-être, la signature de planètes de type terrestre propices au développement d'une chimie de la vie.

Les travaux présentés dans cette thèse se proposaient deux objectifs :

1. La détermination d'un jeu réduit de formes pour les ouvertures photométriques de la voie exoplanètes de Corot , répondant aux multiples contraintes du logiciel de bord tout en évitant le plus possible de dégrader la photométrie.
2. L'élaboration d'une méthode de détection des transits planétaires dans les courbes de lumière qui seront produites en fenêtrant à bord l'image d'un champ d'étoiles, et permettant d'exploiter au mieux les spécificités de Corot .

10.1 Acquisition des données

Nous avons montré qu'un nombre limité d'ouvertures était suffisant pour assurer la photométrie optimisée d'un grand nombre d'étoiles. 250 "patrons" doivent ainsi permettre de fenêtrer les 12 000 étoiles cibles d'un champ sur le CCD de Corot . Une méthode spécifique, dite de "réduction", pour obtenir ces patrons a été développée.

A chaque étoile cible est associée une fenêtre photométrique dont la forme est choisie parmi les patrons. Malgré la perte de spécialisation par rapport au cas

idéal où la photométrie de chaque étoile serait réalisée avec un fenêtre dessinée sur mesure, le S/B ne subit pas de dégradation notable.

Cette méthode de réduction opère sur une collection d'étoiles (i.e) un ensemble d'étoiles extraites d'une série d'images simulées représentatives de la diversité des champs d'observation de Corot . Elle présente 5 étapes différentes :

1. Détermination du "masque optimal" de chaque étoile, c'est-à-dire l'ouverture qui maximise son S/B en fonction de sa position, de son type spectral et des étoiles qui peuplent l'arrière-plan.
2. Mesure de la robustesse de chaque étoile, en termes de S/B , envers les masques optimaux d'autres étoiles de la collection choisies par tirage aléatoire. Notre procédure exploite le fait que les étoiles se montrent tolérantes à des degrés divers envers d'autres masques que le leur.
3. Définition d'un critère d'"acceptabilité" pour chaque étoile cible, sous forme du niveau de préservation de son S/B compatible avec l'objectif scientifique de Corot .
4. Tri optimisé des étoiles cibles et des masques sélectionnés lors du tirage, cibles et masques étant disposés dans une matrice binaire appelée "matrice d'acceptabilité".
5. Ajustement du seuil pour obtenir 250 patrons au plus.

Enfin, partant de données d'observation d'un champ d'étoiles réel, nous avons mis au point une procédure de distribution de ces patrons à toutes les cibles potentielles. Cette procédure utilise une méthode itérative de "repêchage" qui permet de récupérer jusqu'à 50% des cibles qui avec d'autres méthodes étaient considérées perdues pour la mesure à cause des "collisions" entre fenêtres.

L'objectif fixé pour le fenêtrage du CCD (6 000 cibles avec 250 patrons) est atteint puisque seules 10% des étoiles perdent plus de la moitié de leur S/B idéal. De surcroît ces étoiles s'avèrent les moins "intéressantes" pour les objectifs de Corot , grâce à un mécanisme de gestion des priorités combinant le critère de S/B et la priorité scientifique des cibles. On constate que $\geqslant 80\%$ des étoiles conservent leur S/B idéal.

Au-delà du problème abordé dans cette thèse se pose la question du maintien de ces résultats quand les images de travail simulées seront remplacées par des images réelles, plus fiables mais moins riches en informations sur la source des photons d'un pixel. Nous avons 2 pistes pour limiter l'effet du manque de précision dans les PSF :

1. Les étoiles faibles, dont les PSF sont les moins bien déterminées à partir de l'image, sont en revanche les plus tolérantes envers l'erreur sur le patron affecté.
2. Certaines composante du S/B seront néanmoins connues avec une meilleure précision que les PSFs. Il s'agit du bruit photonique, de "jitter" et de "respiration" qui se calculent à partir du flux total accessible par l'image. L'imprécision sur le S/B calculé devrait s'en trouver amoindrie.

10.2 Détection des transits

Nous avons montré le bénéfice, en termes de détectivité, qui peut être tiré d'une approche collective du débruitage des courbes de lumière. L'approche collective s'avère aussi très utile pour gérer de manière automatique et rationnelle les artefacts imprévus sans avoir besoin d'un modèle spécifique bâti a priori. C'est certainement là que réside son principal avantage.

L'originalité de cette méthode est aussi de placer l'étape de détection avant celle de débruitage, contrairement aux méthodes standard. L'interversion de ces deux étapes a pour avantage de rehausser le contraste des signaux de transit en utilisant l'information la mieux connue (la "forme" des transits) avant que celle-ci ne soit altérée par les processus de filtrage.

C'est en travaillant sur 1 000 courbes de lumière simulées pour un test en aveugle dans le cadre de la mission Corot , et confronté à la prépondérance des motifs communs participants au bruit que j'ai été conduit à envisager une stratégie collective. Cette approche a pour but de regrouper l'information éparse afin de pouvoir mieux l'étudier et la corriger. Cette approche met à profit l'avantage de disposer d'un grand nombre de courbes de lumière acquises de manière synchronisée, dans des conditions de mesures stables et sur de longues périodes comme cela est le cas avec une expérience de photométrie dans l'espace telle que Corot .

La méthode que nous avons développée pour traiter les courbes de lumière comprend trois étapes :

1. Une détection à l'aide d'une corrélation glissante entre une section de courbe et un transit de référence nommé "gabarit". Ce détecteur est insensible aux facteurs d'échelle, tient compte du bruit local et tolère les données manquantes sans utiliser d'interpolation. Des essais de robustesse ont montré que dans la pratique deux gabarits étaient suffisants.
2. Un débruitage qui identifie les bruits systématiques ayant survécu à la détection et utilise une analyse en composantes principales et un classement

des bruits par familles. Seuls sont conservés les pics de détection qui se démarquent significativement de la tendance commune.

3. La périodicité des événements détectés est contrôlée, et ceux-ci sont caractérisés en repliant les courbes brutes après recentrages individuels. L'examen séparé des occurrences paires et impaires aide à distinguer les éclipses dissymétriques des étoiles de fond binaires.

Le principal avantage de notre méthode est de traiter automatiquement et de façon pondérée les effets déterministes (identifiés ou non) ou certaines de leurs combinaisons imprévues donnant lieu à des artefacts, qui sinon sont considérés comme du bruit, voire des données perdues.

Elle a permis de réduire de façon significative le niveau de bruit dans les courbes de détections en y éliminant les déterminismes cachés. Par exemple la principale fausse détection présente dans toutes les courbes, due à l'effet imprévu de la conjonction entre la SAA et le pic de lumière diurne, et qui gênait la plupart des autres méthodes du test, a disparu d'elle-même ; les données correspondantes ont pu être conservées contrairement au cas général. Cette efficacité dans l'identification est aussi illustrée par la mise en évidence involontaire de quelques-uns des profils de variabilité stellaire simulée qui avaient été dissimulés à l'intérieur des courbes de lumière du test en aveugle.

Les transits détectés s'inscrivent dans la moyenne des autres méthodes ce qui, compte tenu de la performance modeste du détecteur confirme l'intérêt d'une méthode de débruitage collectif.

A l'exemple du pic commun, il est probable que la plupart des méthodes aient conservé dans leur bruit de détection d'autres déterminismes moins visibles que nous avons mis à jour. La prospection et la suppression de ceux-ci ferait certainement progresser le seuil de détectivité attendu aujourd'hui pour Corot .

L'intérêt de traiter statistiquement les effets collectifs pour améliorer la détection des transits ne fait aujourd'hui plus de doute. D'autres méthodes l'ont confirmé en pratique sur des données obtenues à partir du sol. L'une d'elles pousse d'ailleurs plus loin l'efficacité de cette stratégie en lui associant l'utilisation de modèles.

Loin d'être en concurrence, toutes les méthodes de débruitage et de détection sont complémentaires. Les méthodes statistiques éliminent les perturbations répertoriées, non répertoriées ou fugitives. La nôtre se spécialise dans un "état des lieux" préventif permettant de guider le choix des modèles et de traiter les résidus ; les méthodes de filtrage quant à elles traitent efficacement la varia-

bilité stellaire ; les méthodes de détection par repliement combinent au mieux la forme et la période des transits. Seule une association et des perfectionnements à ces techniques permettront de détecter des planètes toujours plus petites.

Beaucoup d'améliorations sont envisageables au rang desquelles l'utilisation d'une stratégie semblable pour l'identification et la caractérisation statistique des types de variabilité stellaire considérées comme des "textures" de bruit.

Annexe A

Performances des patrons

La table A.1 montre l'influence des paramètres de jitter et de respiration sur les patrons et les affectations. Les patrons ont été réduits avec différents coefficients, exprimés en pixels, utilisés dans le calcul du S/B . Les étoiles affectées dans les mêmes conditions. C'est la respiration (L9111) qui provoque la plus grosse perte de performances.

TAB. A.1 – *Etoiles affectées et S/B total en fonction des patrons. Les patrons ont été obtenus avec diverses valeurs de jitter et de breathing. Les colonnes sont 1) le nom de la collection de patrons, 2) et 3) les conditions de jitter et de respiration qui prévalaient lors de la réduction, 4) le nombre d'étoiles affectées qui gardent S/B* $\geqslant 100$ *(sur 8.5 min), 5) le nombre d'étoiles affectées total, 6) le S/B total pour les étoiles de la première catégorie, 7) le S/B total des étoiles affectées et 8) le S/B potentiel si toutes les étoiles étaient affectées sans perte.*

1 nom	2 jitter	3 respiration	4 n 100	5 n total	6 S/B 100 ($(\times 10^4)$)	7 S/B total ($\times 10^4$)	8 S/B optimal ($\times 10^4$)
L6111	0.03	0.00	5106	5818	116	122	123
L5111	0.10	0.00	4690	5784	108	116	118
L7111	0.20	0.00	4237	5829	98	109	112
L2111	0.35	0.00	3651	5729	85	98	101
L9111	0.40	0.04	1447	5614	28	49	50
L8111	0.50	0.00	3368	5718	78	91	94

Annexe B

Le test en aveugle

B.1 Article de Moutou et al. (2005), A&A.

Astronomy & Astrophysics manuscript no. paper˙blind˙final February 21, 2005
(DOI: will be inserted by hand later)

Comparative blind test of five planetary transit detection algorithms on realistic synthetic light curves

Moutou, C.[1], Pont, F.[1,5], Barge, P.[1], Aigrain, S.[2], Auvergne, M.[3], Blouin, D.[1], Cautain, R.[1], Erikson, A. R.[6], Guis, V.[1], Guterman, P.[1,7], Irwin, M.[2], Lanza, A. F.[4], Queloz, D.[5], Rauer, H.[6], Voss, H.[6], Zucker, S.[5,8]

[1] LAM, Traverse du Siphon, BP8, Les Trois Lucs, 13376 Marseille cedex 12, France
[2] Institute of Astronomy (IoA), University of Cambridge, Madingley Road, Cambridge CB3 0HA, United Kingdom
[3] OPM, Place J. Janssen, 92195 Meudon cedex, France
[4] INAF-Osservatorio Astrofisico di Catania, Via S. Sofia, 78, 95123 Catania, Italy
[5] Observatoire de Genève, 51 Chemin des Maillettes, 1290 Sauverny, Switzerland
[6] DLR Institute of Planetary Research, Rutherfordstr. 2, D-12489 Berlin, Germany
[7] Gemplus Card International, La Ciotat, France
[8] Present adress: Faculty of Physics, Weizman Institute of Science, Rehovot 76100, Israel

Received date / accepted date

Abstract. Photometric surveys for exoplanet transits are very promising sources of new discoveries for future years. Consequently, many algorithms are being developed to detect transit signals in stellar light curves. This paper discusses the comparison of such algorithms for the next generation of transit detection surveys: dedicated space missions like CoRoT, Kepler and Eddington. The comparison of five independent analyses of a thousand synthetic light curves is presented here. The light curves were produced with an end-to-end instrument simulator and include stellar micro-variability and a varied sample of stellar and planetary transits diluted among a much larger sample of light curves. The results show that different algorithms perform quite differently, with varying degree of success in detecting real transits and avoiding false positives. We also find that the detection algorithm alone does not make all the difference, the way the light curves are filtered and detrended beforehand also has a strong impact on the detection limit and on the false alarm rate. The microvariability of sun-like stars is a limiting factor only in extreme cases, when the fluctuation amplitudes are large and the star is faint. In the majority of cases it does not prevent the detection of planetary transits. The most sensitive analysis is performed with periodic box-shaped detection filters. False positives are method-dependent, which should allow to reduce their detection rate in real surveys. Background eclipsing binaries are wrongly identified as planetary transits in most cases, a result which confirms that contamination by background stars is the main limiting factor. With parameters simulating the CoRoT mission, our detection test indicates that the smallest detectable planet radius is of the order of 2 Earth radii for a 10-day orbital period planet around a K0 dwarf.

Key words. Planetary systems - Methods: data analysis - Techniques: photometric - Stars: activity

1. Introduction

Transit searches have recently shown their potential in discovering planetary candidates. The ground-based OGLE project, for instance, (Udalski et al. 2002a,b, 2003, 2004) detected 177 planetary transit candidates, among which so far 5 are confirmed as short-period planets (Konacki et al. 2003; Bouchy et al. 2004; Pont et al. 2004; Konacki et al. 2005, submitted). Space-based transit searches are expected to be much more efficient, because of i) their continuous time sampling over long periods, ii) the more stable photometric signal. At precisions of a few mmag, the main limitation comes from residual systematics due to the instrument and from intrinsic stellar variability. These are the problems that the transit detection algorithms should face in future space missions for long-term planet searches: CoRoT (Baglin 2003), Kepler (Borucki et al. 2004) and Eddington (Favata 2004).

Several transit detection algorithms were proposed in the recent literature: Bayesian algorithms (Doyle et al. 2000; Defaÿ et al. 2001; Aigrain & Favata 2002), matched filters (Jenkins et al. 1996), box-shaped transit finder (Aigrain & Irwin 2004) and the Box-fitting Least Squares (BLS) method (Kovács et al. 2002). A theoretical compar-

Send offprint requests to: Claire.Moutou@oamp.fr

ison of these methods was proposed (Tingley 2003) which concluded that "no detector is clearly superior for all transit signal energies", but an optimized BLS algorithm still performs slightly better for shallower transits. Here, we adopt a more empirical approach to make the comparison: we use as a testbench synthetic light curves with detailed simulations of the instrumental noise and astrophysical sources of variability, to blindly test five different transit detection techniques. The five different detection teams have no prior knowledge on their content.

This comparison of detection algorithms is likely relevant for all transit-search programmes, from the ground and from space, although it has been focussed here on CoRoT, the first space mission largely dedicated to transit searches, to be launched in 2006. The CoRoT characteristics are given in Boisnard & Auvergne (2004) and its planet detection capability is estimated in Bordé et al. (2003). This ability is empirically addressed in this paper.

The goals of this blind detection simulation are the following:

- To independently apply several light curve analysis methods on the same simulated light curves, removing the possible "subjective" elements (possible biases when the same person simulates the transit and detects it)
- To compare their ability to detect faint transits, avoiding false positives (false positives beeing hereafter defined as the noise features from instrumental or stellar micro-variability origin, accidentally picked up as a transit signature).
- To estimate the impact of star micro-variability for transit searches.
- To test the ability to distinguish between a planetary transit and an eclipsing binary from the light curve alone.

Applied to CoRoT, this exercise will help deriving an estimate of the detection limits of this instrument and its limiting factors, as well as defining the strategy for light curve analysis and required follow-up.

Section 2 presents the light curve building procedure; Section 3 then describes the five light curve analysis methods and Section 4 discusses the results and draws conclusions.

2. Generating simulated light curves

The synthetic light curves were built by combining several components: the instrumental model, stellar micro-variability, and in some cases a planetary transit, eclipsing binary or variable star signal.

2.1. Instrumental model

An instrument model (Auvergne et al. 2003) has been designed for CoRoT in order to evaluate the instrument detection capabilities and test the onboard and ground-based software. We use the output of this model as the basis of our synthetic light curve construction. Let us recall that the CoRoT onboard software will perform photometry on a pre-determined list of stars (12,000 per pointing) every 8 minutes during 150 days, by summing all the signal within pre-defined aperture covering between 100 and 60 pixels depending on the magnitude. Environmental perturbations such as light scattered by the Earth, radiation flux, Attitude Control System jitter and temperature variations are computed by specialised models. The outputs are light curves at the focal plane level, proton fluxes with a 10 mm CCD shielding, satellite angular depointing and temperature curves for the most sensitive sub-systems. Monochromatic PSFs are then provided using an optical model of the telescope, and used to compute white PSFs, taking into account the optical transmission, CCD quantum efficiency and target flux for main sequence stars in the effective temperature range 3500 to 9000 K. The appropriate photometric aperture is computed, depending on the star position, magnitude and colour (Llebaria et al. 2003).

We build 25 basic light curves based on stars scanning 5 magnitudes, from 12 to 16, and 5 temperatures, from 4500K to 6750K, all located at the same CCD position. They contain the following realistic noise contributions:
(i) Photon noise (Poisson statistics).
(ii) Flat-field noise, with a 1% non-uniformity.
(iii) Read-out noise of 10 electrons/pixel/read-out.
(iv) No jitter amplitude; it is negligible in the CoRoT broad bandpass.
(v) Zodiacal light, the unique source of sky background in space, a uniform offset of 12 electrons/pixel/second over the CCD remaining constant along the orbit. It is corrected by subtraction and the resulting additional photon noise is kept.
(vi) Proton impacts. The exposures corresponding to the crossing of the South-Atlantic (SAA) anomaly are not usable and the final data thus contain a large number of quasi-periodic gaps (typical duration of 30 min each interval of 1.7 hours) that should be handled by the detection algorithms (Figure 1).
(vii) Earth scattered light, which is is not uniform over the CCD, and varies along the orbit, almost following the orbital period. We insert a scattered light contribution with a realistic maximum value of 1 electron/pixel/second. As it will be corrected in the processed CoRoT light curves to a certain level, we subsequently remove the scattered light contribution to first order, leaving a random $< 50\%$ residual. The correction applied may lead to a positive or a negative residual signal, corresponding respectively to an overestimation or underestimation of the actual scattered light level (Figure 1). This allows (i) to test the robustness of the detection algorithms, especially against a negative (i.e. when it is over-corrected), quasi-periodic signal, and (ii) to create 999 light curves with varying scattered-light noise amplitudes, produced from a parent set of 25 instrumental curves. Note that scattered light is the dominant systematic signal in the CoRoT instrumental noise and the only instrumental systematics included in the simulation;

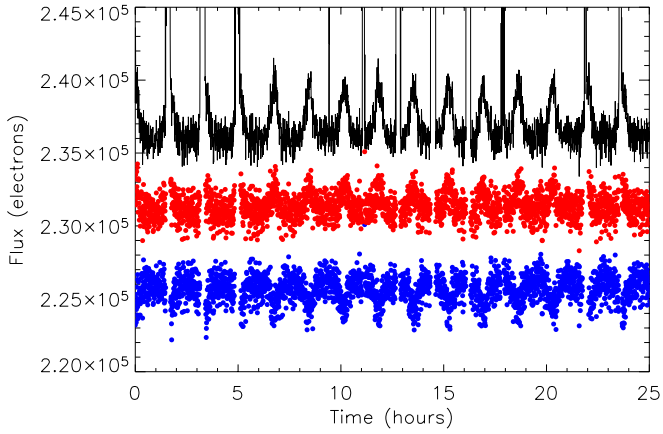


Fig. 1. Example of an instrumental light curve before (top) and after the partial correction of scattered light (once underestimated (middle), and once overestimated (bottom)). The sharp peaks in the upper plot are due to the SAA crossing; they become gaps in the output light curves.

this is the reason why we deliberately took a conservative value for its level of correction.

2.2. Stellar micro-variability

Stellar micro-variability curves are taken from two independent models. These effects are independent of the instrument and are usually thought to be among the main limitations of transit detection. Considering two types of micro-variability curves, there are 55 different light curves. To build the final light curves the micro-variability curves are all scaled by a random factor between 0.5 and 2, to account for the dispersion in the variability level observed in real stars. They are also phase shifted by an arbitrary value, and rebinned in their time sampling by a random factor between 1.0 and 1.2, to avoid excessive similarity between the final light curves.

2.2.1. A scaled solar-like model for stellar variability

Lanza et al. (2003) model the variation of the Total Solar Irradiance (TSI) by considering a simple stellar-like model based on the rotational modulation of the visibility of three active regions plus a uniformly distributed background component which accounts for the surface features affecting the mean level of the solar flux. Each active region consists of faculae and sunspots in a fixed area ratio and with brightness contrasts that are functions of the limb angle. A time interval of 153 days ranging from 1^{st} July to 1^{st} December 2000 is selected as representative of the variability of the TSI close to the maximum of the 11-yr cycle. The model is applied to successive subintervals of length 14 days, separated by 7 days from each other, to obtain the coordinates and the areas of the three model active regions plus the uniform background term.

In order to simulate optical light curves for main-sequence stars rotating faster than the Sun and with a higher activity level, the rotation period and the areas of the three model active regions are varied: the areas of the three active regions as well as the uniform background term are multiplied by a factor $f = A(P, Sp)/A_{\odot}$, where $A(P, Sp)$ is the average amplitude of the optical light curves of a star of rotation period P and spectral class Sp derived from Messina et al. (2003) and $A_{\odot} = 2.2 \times 10^{-3}$ mag is the maximum amplitude of the solar optical variability. For stars with a rotation period longer than 12 days, there is no information on the amplitude of the rotational modulation in the optical passband (except for the Sun), so that f is assumed to be in the range 1.5 to 6 for a spectral type varying from F5V to K5V. The coordinates of the three active regions are those of the solar model active regions and the inclination of the stellar rotation axis with respect to the line of sight is fixed at 90°. To reduce the impact of the small discontinuities occurring every 7.0 days at the passage from a fit to the next, the model parameters are linearly interpolated in time between successive best fits. The brightness contrast coefficients and their center-to-limb variations are the solar ones. The ratio of the area of the faculae to that of the sunspots in an active region is estimated by extrapolating the relationship given by Chapman et al. (1997) to larger sunspot areas. The resulting facular contribution is found to be negligible for stars with a rotation period shorter than 20 days and spectral type later than G8. The variability on time scales significantly shorter than the rotation period is modelled by scaling the residuals of the best fits to the solar TSI variations, which are due to the evolution of the solar active regions on time scales shorter than 4-5 days (Lanza et al. 2003, 2004). In order to increase the amplitude of the short-term stellar variability to make the planetary transit search more challenging, the residual solar variability is multiplied by a factor $3f$ and linearly interpolated to get an even time sampling of 8 minutes. Finally, Poisson random fluctuations with a relative standard deviation of $[3 \times (fA_{\odot})^2]^{-0.5} = 3.8 \times 10^{-3} f^{-1}$ are added to simulate short-term variations due to microflaring or convection on time scales of several minutes.

In addition to the original TSI light curve, 9 light curves were produced with this method, with spectral types F5, G0 and G8 and rotation periods 3, 10 and 20 days. The amplitude of micro-variability ranges from 0.1 to 4 %. The stellar optical time series so obtained are dominated by the rotational modulation except for rotation periods longer than 15-20 days for which the active region evolution prevails on the rotational modulation signal. A few small discontinuities are present, due to the passage from a 14-d fit to the successive one, but they never exceed 5% of the amplitude of the rotational modulation, even in the case of the most active stars.

2.2.2. Light curves from SIMLC

SIMLC is a tool to simulate stellar micro-variability for stars with spectral types F5 to K5 and ages later than 625 Myr. It works by computing an artificial power spectrum, starting from a fit to solar data and scaling it using empirical scaling laws. The power spectrum is then sampled as appropriate given the time sampling and light curve duration required, coupled with a random phase array, and reverse Fourier-transformed to the time domain. More details can be found in Aigrain et al. (2004), and only a brief summary is given here.

Following Andersen et al. (1994), the power spectrum of the Sun's total irradiance variations up to $\sim 600\,\mu$Hz (as observed with the PMO6 radiometer, which is part of the VIRGO experiment on SOHO), is modelled as a sum of three broken power laws, each characterised by an amplitude, characteristic timescale and slope. There are 3 components, with timescales of 10 days, 4 days and 10 min. The powerlaw slopes are 3.8, 1.8 and 2.0. All these values are those measured for the Sun. Note that because of the slope of the first powerlaw is quite steep it falls of quickly for timescales larger than 10 days, while the second powerlaw, which is quite shallow, is still the dominant component at 100 μHz (timescales of a few hours, typical of transits). The amplitude of the lowest frequency, or 'active regions', component is correlated with simultaneous measurements of the CaII K-line index indicator of chromospheric activity. Higher frequency components, which have much smaller amplitude, are thought to be related, respectively, to super- or meso- granulation and to a superposition of granulation, oscillations and photon noise.

Empirically derived scaling laws can be used to scale the amplitude and timescale of each power law to what might be expected for other stars. Currently this can be done only for the dominant low-frequency component, using chromospheric activity as a proxy. Observational constraints are currently insufficient to derive scaling laws for the other components, including the second component which corresponds to the timescales characteristic of planetary transits, and those are thus left as they are measured in the Sun. Upcoming data, in particular from the MOST (Micro-variability and Oscillations of STars) satellite (Walker et al. 2003), are expected to provide constraints on this component in the near future.

A set of 45 light curves lasting 150 days, with 8 min sampling, were generated for the present exercise. They correspond to a grid of stars of spectral type F5, F8, G0, G2, G5, G8, K0, K2 & K5, and ages 0.625, 1, 2, 3 and 4.5 Gyr. The amplitude of the dominant, 'active regions' component of the variations scales with convection zone thickness (which is larger in later spectral types) and the inverse of the rotation period (which is larger in older stars), while the characteristic timescale scales roughly with the rotation period. As a result, at 0.625 Gyr the most variable stars are F-stars, while at 4.5 Gyr they are K-stars. The amplitude of micro-variability ranges from 0.01 to 0.1%, a level much lower than those obtained with the method described in Section 2.2.1. This is thought to be due to the more coherent nature of micro-variability in active stars, which SIMLC currently cannot reproduce.

2.3. Transits

Twenty planet transits were simulated. For a thousand light curves, this represents about an order of magnitude more transit events than expected in real samples (Bordé et al. 2003). It is important that light curves without transit vastly outnumber those with transits in the simulation, so that the detection thresholds have to be set realistically high. The characteristics of the inserted transits are not chosen with the goal of reproducing planet statistics, because those are mostly unknown in the range where CoRoT will discover planets; the idea is rather to test limitations and to explore the borders of detectability. The objectives are then (1) to sample a variety of system cases and (2) to investigate the detection limit by including a large number of small planets in light curves with a varying noise level. The characteristics of the transits are summarized in Table 1. The planet size spans the range from 1.6 Earth radius (R_E) to 1.3 Jupiter radius (R_J). One system with two planets is inserted. The period domain is 4 to 90 days. The target stars with the planetary transits are chosen at "directed random", with the aim of exploring the regions near the limit of detectability. For instance, the largest planets are inserted in the light curve of faint and/or active stars. The largest planets are also the ones with the lower number of transits (the hot Jupiter configurations, beeing easy cases for space transit searches, are not emphasized here).
The transit light curves are simulated with the aid of the Universal Transit Modeler (Deeg 1999). Limb darkening of stars are estimated from recent calculations from ATLAS9 models and the CoRoT bandpasses (Barban, priv. comm., see method in Barban et al. (2003)), taking into account a linear limb-darkening law and a classical mixing-length theory.

2.4. Eclipsing binaries and large-amplitude variable stars

Simulations (Brown 2003) and the results of the OGLE planetary transit follow-up (Bouchy et al. 2004b, Pont et al. 2004b) indicate that for a given transit signal depth, the contamination by grazing and background eclipsing binaries (EB) will be at least as numerous as the planet transits themselves, or could even largely outweigh the true planet events. To simulate this contamination, we inserted ten low-depth stellar eclipse signals among the light curves. There are grazing binaries (6 events), background binaries (4 events) and one hierarchical triple stellar system. Finally, we inserted five background variable stars: a low-amplitude delta Scuti, a classical Cepheid, a β Cephei, the semi-regular variable Z UMa and the irregular Z Cam. The background variables and background eclipsing bina-

ries refer to fainter objects included in the same aperture, 3 to 7 magnitudes fainter than the main target. In the case of grazing eclipses, the binary star is the main target itself, i.e. in the magnitude range 12–16 for CoRoT.
The characteristics of these light curves are summarized in Table 2.Again, the characteristics of the systems are chosen to cover most possible combinations rather than to reproduce the expected characteristics of real samples. Our eclipsing binary transits include curves with anti-transit signals, with sine and double-sine modulations outside the transits due to the ellipsoidal deformation of the primary under the gravitational influence of the secondary, V-shaped eclipses (grazing) and U-shaped eclipses (central eclipse in a background contaminant system). For grazing eclipsing binaries, the algorithms of Mandel & Agol (2002) and Wichmann (1998) are used. The Universal Transit Modeler (Deeg 1999) is used for background eclipsing binaries and the triple star. The variable star light curves are taken from the literature and from the archives of the AAVSO (American Association of Variable Star Observers).

2.5. Crowding

Another consequence of background stars is to contribute to the flux variations measured in the aperture placed on the primary target. To simulate this effect, we systematically added to the primary light curve the contribution of one background star, characterised by a light curve constructed with the same procedure as for the main target, and a magnitude difference with a distribution probability $\sim 2^{\Delta m}$ in the range 0-6 magnitude (thus including stars up to 22^{th} magnitude). For the second star, another stellar micro-variability curve is used. Thus, each final simulated light curve consists in the addition of two different contributors.

2.6. The final set of 999 simulated light curves

The sample of 999 light curves was composed from a combination of the individual elements described so far, as it is developed in this section. The parent lightcurves are the 25 instrumental curves (from a grid of 5 magnitudes and 5 color temperatures), with a level of scattered-light residual noise different in all lightcurves. The magnitude of each target was drawn from a probability distribution: $p(m) \sim 2^m$, approximating an isotropic distribution near the Galactic plane, between 12 and 16 mag. The distribution of color temperatures was selected to roughly match a spectral type distribution realistic for magnitude-limited, transit-search fields near the Galactic plane; from an analysis of the stellar population in future CoRoT fields, there are 40% of F dwarfs, 40% of G dwarfs and 20% of K dwarfs (Moutou et al, in prep.). Finally, the micro-variability fluctuations were inserted: from the 55 parent light curves, all final micro-variability contributions are unique due to the applied amplitude and temporal extension factors (section 2.2), and due to the injection of a fainter stellar light curve (section 2.5). The micro-variability light curve was also selected to match the color temperature (or spectral type) of the instrumental light curve. In total, 964 light curves do not have any transit or EB/variable star signal.

The temporal sampling of the final light curve is 8 minutes, with a duration of 150 days, as for CoRoT long observing runs. A complete light curve contains 25056 data points.
The package of 999 light curves (identified with ID 1 to 999 in the following) were supplied to the detection teams with no information on their content nor on the way they were calculated; neither the number of hidden planets nor the nature of injected noise sources were known by the detecting teams. In the real case with CoRoT light curves, some data will be known beforehand, such as the star magnitude, spectral type, luminosity class, contamination by neighbours, and pipeline processing parameters. This knowledge is not fundamental for transit detection but will obviously help in the identification of the detected events.

3. Blind search for transit events

In this section, we describe the five methods used for detrending the light curves and detecting the transits. Their elements span a wide range of complexity from fairly basic to very evolved. They also differ by their previous use: one team started from scratch with no experience in transit detection, two teams use algorithms that they developed for ground-based transit surveys (BEST and OGLE), and two teams are working on algorithms for space-based transit searches.

3.1. Team 1: correlation with a sliding transit template

The first algorithm is based on correlation of the light curve with a single sliding template, without prior detrending. Systematic noise on short timescales is removed from the correlation function, then candidates with a high signal in the correlation function are examined individually by eye to pick up the final detections.

Detecting the transits: The light curves are correlated with a sliding template to compute a correlation function C(t). The template is a transit shape based on the algorithm of Mandel & Agol (2002). The use of a unique transit template is sufficient and makes the method much simpler; the optimum template has a transit duration of ∼8 hours and is bordered by two flat segments of ∼14 hours. Previous filtering of the long-term variations is not crucial in this case, because the template covers only a small part of the light curve at a time. Fig. 2 shows the resulting correlation functions for a few cases. In this method, no periodicity is assumed in the transit signal and the period is estimated a posteriori.

One advantage of the correlation method is that it is not affected by gaps in the time coverage of the data.

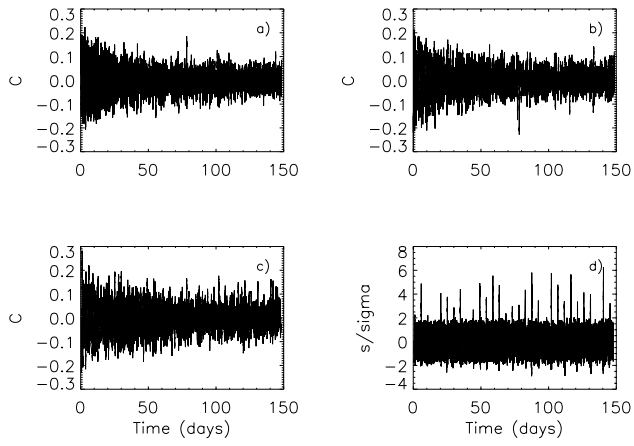


Fig. 2. A) and B) are two correlation functions ("detection curves" DC) showing systematic noise. Artefacts are sometimes obvious (synchroneous spikes and similar envelope) or can be hidden, with a known or unknown origin. C) and D) show DC613 before then after detrending (note the very different y-axis scales).

Missing epochs simply make no contributions to the correlation function, which avoids the problems caused by any interpolation of the data in the gaps.

Detrending the light curves: As explained above, no detrending was done on the long-term variations. Correlation curves show a common pattern of perturbation on short time scales, associated with instrumental effects like temperature changes ("breathing"), scattered light or pointing jitter. We assume that this instrumental noise introduces a common noise in all correlation functions, except for a scale factor. We model this by $\overrightarrow{C_i} = \overrightarrow{s_i} + \lambda_i.\overrightarrow{p}$ where $\overrightarrow{C_i}$ is the temporal correlation curve, $\overrightarrow{s_i}$ is the unknown noise-free correlation curve, p (with $\|\overrightarrow{p}\| = 1$ by convention) the unknown instrumental perturbation common to all objects, weighted by the unknown λ_i. It appears that the average of λ is close to zero, so that p cannot be simply estimated by averaging the curves. To retrieve p we apply the following sequence:

1. Choose an initial guess for p from a light curve strongly dominated by p;
2. Estimate a first-guess λ_i by projecting $\overrightarrow{C_i}$ on $\overrightarrow{p}$;
3. The mean of C/λ over all objects is our refined p, giving the refined λ_i.

Actually, we found that the instrumental noise pattern was not common to all stars, but could be classified into a number of families. We therefore applied the above procedure to determine different p for each empirically determined family.

Detrended correlation functions exhibiting a strong signal (i.e. about 5% of the light curve sample) are then examined by eye, selecting the candidates with strictly periodic signals and folding accordingly each light curve to point out autosimilarity of the shape.

Discussion: It turns out that the "families" of objects used to remove the noise in the correlation function often correspond to sets of light curves based on the same parent noise curve. Therefore, with this method, the removal of the systematic noise is probably more efficient on simulated data than it would be in reality.

Correlation with a sliding transit template is among the simplest possible methods for transit detection, short of direct examination of all light curves by eye, and the results of this algorithm on our synthetic sample can be used as a reference point of comparison for the performances of the other algorithms.

3.2. Team 2: Box search with lowpass filtering and broken-line detrending

The algorithm searches for box-shaped signals in normalized, filtered, variability fitted and unfolded light curves. It was designed to detect single as well as periodic transit events.

Detrending the light curves: In a first step all the light curves are normalized, neglecting all the epochs without flux value. These epochs, covering a maximal time span of 43 minutes caused by crossings of the SAA, are short compared with the transit durations of minimal 2 hours. Therefore the missing epochs are linearly interpolated, without risk of introducing false transit events. A Fourier analysis is carried out, giving a dominant systematic periodic signal at a period of P$\sim$1.13 hours – the orbital period of the satellite (residuals of the scattered light contribution). A standard lowpass-filter is used to eliminate this signal and other high frequency signals. The cut-off frequency is varied between 0.059 day^{-1} and 0.177 day^{-1}. The shape of the transit signals is moderately deformed by this kind of filtering, but for the purpose of a detection tool the influence of this side-effect is negligible. Another side-effect of the lowpass-filtering is that an additional modulation of the light curves occurs at the beginning and the end of the data. Therefore the data of the first and last 10 days are excluded from the transit search. The stellar variability is fitted locally. The light curves are separated into sub-sections and a linear least-squares fit to the data is performed for every sub-section. The size of the sub-sections is varied in a range between 0.5 to 3 days to ensure that no transit-like signals are significantly altered. In a subsequent step the fit is subtracted from the data (see Fig. 3).

Detecting the transits: The standard deviation of the normalized, lowpass-filtered and variability-fitted light curves is calculated. Subsequently a box search for transit-like events is carried out. All data points deviating from the

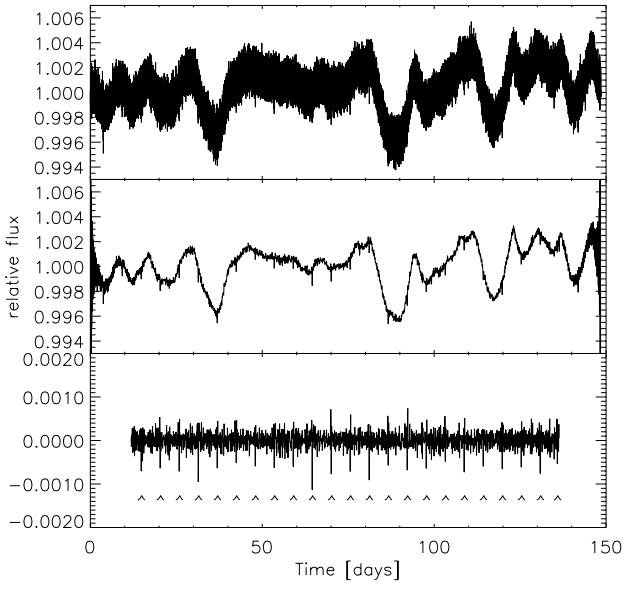


Fig. 3. In the figure the different steps of the light curve analysis of team 2 is exemplified for light curve #34. First the data gaps (not visible at this scale) are interpolated and the light curve is normalized (top) and a lowpass-filter is applied to remove high frequency signals (middle). Finally the stellar variability is modeled and a search for period signals is performed (bottom). The periodic signal found is marked in the figure.

average signal by 3σ are identified and the neighbouring deviating points are combined into a single detection. A maximal and a minimal signal length are defined, corresponding to transit lengths between 1 to 30 hours. Mean epochs of the signals are determined, spurious detections are excluded (this mostly concerns an instrumental artefact that is identified in all light curves) and all remaining single detections are listed for further inspections. Thereafter the epochs of the potential events found are automatically searched for periodicity: time differences between all detected events with approximately the same detection level are estimated and retained when a single time difference or multiples of it have occured several times within a given error margin. For possible cases a detailed investigation of the potential transit events is performed where the depths and duration of the events are determined. The corresponding light curves are manually inspected for secondary eclipses and gravitationally induced modulations caused by high-mass secondaries.

Finally all light curves with detected events are classified as either possible transit-like or other events.

Discussion: The algorithm is based on a search routine for single transit events developed for the ground-based transit search BEST (Rauer et al. 2004). The adapted version can detect both single and periodic transit-like events. It was also adapted to be able to detrend the microvariability that is not an issue for ground-based wide angle searches. In general, to discriminate between real and false transit signals more information about the host star is helpful, especially for non-periodic single transit events.

3.3. Team 3: Box least-squares with 200-harmonics filtering

In this method, the light curves are detrended by fitting 200+5 harmonics, then transits are detected with a box-fitting on the phase-folded signal.

Detrending the light curves: The scattered light periodic effect is verified to have the same period in all light curves, though its shape is varying among them. Therefore it seems plausible to describe it (in each light curve independently) as a sum of a small number of harmonics (5) of a fundamental frequency – $f_S = 14.18\,\mathrm{day}^{-1}$.

$$F_S(t_n) = \sum_{k=1}^{5} a_k \cos(2\pi k f_S t_n) + \sum_{k=1}^{5} b_k \sin(2\pi k f_S t_n) \ .$$

Separately, the long term stellar variability is also modelled as a sum of harmonics. The fundamental frequency used this time is $f_L = \frac{1}{2T}$, where T is the whole duration of the light curve (about 150 days). The number of harmonics to consider for the long-term variability model, N_L, is fixed to 200. Thus, the highest frequency in this model corresponds to a period of 1.5 day. We expect the energy in a planetary transit signal to be mainly concentrated in higher frequencies, because of the relatively sharp ingress and egress. Therefore the harmonics we fit should include only a negligible fraction of the energy in the transit signal.

$$F_L(t_n) = \sum_{k=1}^{N_L} c_k \cos(2\pi k f_L t_n) + \sum_{k=1}^{N_L} d_k \sin(2\pi k f_L t_n) \ .$$

Including also the average level, the model is parameterized by 411 parameters, estimated by a least-square fit. Naïvely, that would involve solving a system of 411 linear equations (known as the normal equations) with the same number of unknowns. This may be prohibitively time-consuming. Therefore, we consider only the times t_n for which a valid measurement existed for all light curves, i.e., times which are guaranteed to lie outside the SAA. This amounts to about two thirds of the original sampling times. Using only those points, which are common to all light curves, allows us to calculate SVD (Singular Value Decomposition) pseudoinverse of the normal equation matrix (Press et al. 1993) and then use the *same* matrix to solve for the 411 coefficients in each light curve separately. As it turns out, the price we pay by using only part of the points is negligible, because of the very good time coverage. After the fit, the derived coefficients are used to model and remove the scattered light and stellar variability from the complete set of points of the light curve.

Detecting the transits: We apply the Box-fitting Least-Squares (BLS) algorithm, presented in Kovács et al.

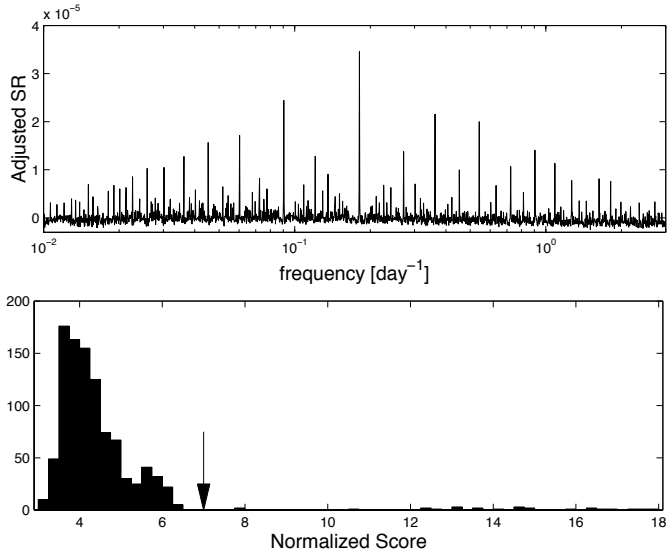


Fig. 4. Top: The adjusted SR function for light curve 34, which shows the typical peaks of a transit signal. **Bottom:** The distribution of the normalized SR of the 999 light curves. The arrow points to the adopted detection threshold of 7.0 (Team 3).

(2002), on the detrended data. We use a logarithmic sampling of the frequency space, with 2000 frequencies between 0.01 and $3\,\mathrm{day}^{-1}$. The maximum allowed transit width is proportional to the cubic root of the period, as suitable for Keplerian orbits, and 5 phase bins are allowed in a maximum width transit. Eventually, a simple function ($a + \frac{b}{f}$, where f is the frequency) is fitted and subtracted from the SR (Signal Residue) function (Kovács et al. 2002) to adjust it for the varying number of configurations tested in each frequency (Fig. 4 top). Fig. 4 (bottom) displays the distribution of the normalized relative heights of the SR peak, for all 999 light curves. One can clearly see a bell-shape distribution, whose samples all lie below a value of 7.0. Thus, we fix 7.0 as the detection threshold, tagging all the scores above it as detections.

Discussion The main attractive feature of the harmonic-fitting procedure is the fact that it does not require any interpolation of the measurements onto a regular grid. Such interpolation would have introduced interpolation noise with some periodic nature, due to the SAA gaps, and probably introduce more false alarms.

Removing harmonics with periods as short as 1.5 days may modify a little the shape of the transit signal, but it does not affect the detection capability. The final characterization of the transit signal is done by fitting, together with the harmonics, a simplified transit model, with linear ingress and egress, and a 'flat-bottom' transit. Fitting it together with the harmonics proved quite easy, using the SVD pseudoinverse method, and the derived transit signal is not modulated by the harmonics.

3.4. Team 4: Matched filter with image-processing detrending

In this method, the signal is denoised with the "Gauging Filter" (GF), and the detection is performed with a standard matched-filter associated to a three-criteria decision process (Guis & Barge 2004).

Detrending the light curves: In order to remove from the signal the low frequency modulations, the GF is applied to the light curves following a procedure described in Guis & Barge (2004). In this method the time plot drawing of a light curve is considered as a 2D-image which splits into two parts: one corresponds to the area below the time plot, the other corresponds to the area above. Then, a denoised signal S_F is defined as the mean value of the two curves resulting from the processing of these two images. The result may still contain residual components at various frequencies. These residuals are removed thanks to a second-order filtering at various scales and Fourier transforms.
The GF detrending procedure is the following: (i) the light curve is successively under-sampled and expanded with a linear recursive interpolation method over the data gaps in order to keep the total size of the light curve unchanged; (ii) the resulting light curve is smoothed out with a 4-width smoothing filter (widths are 2^n with n = 6,7,8,9), producing smoothed light curves with different low frequency ranges; (iii) the final light curve is chosen as the optimum of the four filtered light curves. The final choice is made in Fourier space looking at the local minima of the energy contained in the four light curves and selecting the one within the lowest frequency range (*i.e.* the furthest from the transit frequencies).
For a given light curve, the best fit of the low frequency modulations obtained with our detrending method is denoted C_{LF}. In most light curves, the low frequency modulations are quite weak and using under-sampled light curves with loose smoothing is sufficient.

Detecting the transits: The detection method is based on the classical matched filter aimed at detecting a single feature in a noisy signal (Defaÿ 2001). It is composed of three main steps: (i) a subtraction of C_{LF} from the light curve; (ii) a convolution of the detrended light curve with a reference filter (based on a model of planetary transit) resulting in a convolution curve M; (iii) the identification of local convolution maxima in M which directly provides possible positions for the transit like features. The convolution maxima are selected in two different ways: one is a correlation in Fourier space with a library of periodic signals; the other is a sorting of the convolution peaks and a selection according to criteria based on statistical parameters like the variance or the entropy of the peak distribution (Guis & Barge 2004). In summary, our detection method combines the standard matched filter, which is well suited

for localising isolated features, and the Fourier correlation, which permits to find periodicities more efficiently.

With the above method a total of 25 light curves were found to contain transit-like features (Tables 1, 2). 19 of them are identified thanks to Fourier correlation; 13 (resp. 2) of them corresponds to single (resp. bi-) periodic features present all along the light curve, and 4 have the characteristics of an eclipsing binary. Bi-periodic events are characterized by two not-commensurable periods. Selection by peak sorting allowed to identify the 6 other detections, with a lower confidence level, but also some secondary features.

Discussion: Detection is made using pairs of peaks chosen among the set of the selected peaks. The period corresponding to this pair is then checked against the positions of other selected peaks. The larger the number of the pairs or the shorter the periodicities, the higher the confidence level. In some cases (IDs 983, 985) the noise level is so strong, probably due to star micro-variability, that detections become less reliable. Indeed, the matched filter is very sensitive to strong discontinuities in the signal. Further, it can be noted that selection by peak sorting permits to bring out some potentially interesting cases, as for example signals with rough periodicity on parts of a light curve. However, such cases were removed from our list of possible events because their periodicity was not firmly established.

3.5. Team 5: Box maximum-likelihood with iterative 1-D filtering

Detrending the light curves: Residual scattered light variations, whose period is determined by sine-fitting over the range 0.065 to 0.075 day, are removed by phase-folding each light curve at the best-fit period, smoothing it using a 1–D filter (median, then boxcar filter, with respective widths of 511 & 11 data points), and subtracting the smoothed light curve from the original. Other 'glitches' common to all light curves are removed by scaling each light curve to unit median, computing the median of all scaled light curves, and subtracting a scaled version of this 'common component' from each light curve. Bad data points (large scatter in 'common component' light curve) are also flagged at this stage.

Long term (stellar) variations are then removed using an iterative clipped non-linear filter (Aigrain & Irwin 2004). First the light curve is pre-filtered with a combined median/boxcar filter (duration 7,3 samples) to remove short duration glitches and to minimise the removal of signal from transit-like features. A "continuum" is then computed from this pre-filtered curve by iteratively applying a similar median/boxcar filter (duration 2d,d samples, where d is the trial transit duration), flagging points where the difference between continuum and original is $> 3\sigma$, and recomputing the continuum without the flagged points up to 5 times. The σ is robustly re-computed at each iteration from the median of the absolute deviations of the difference signal. The final clipped continuum is subtracted from the original signal and the median level restored to give the filtered (white-noise-like) light curve (see example in Fig. 5).

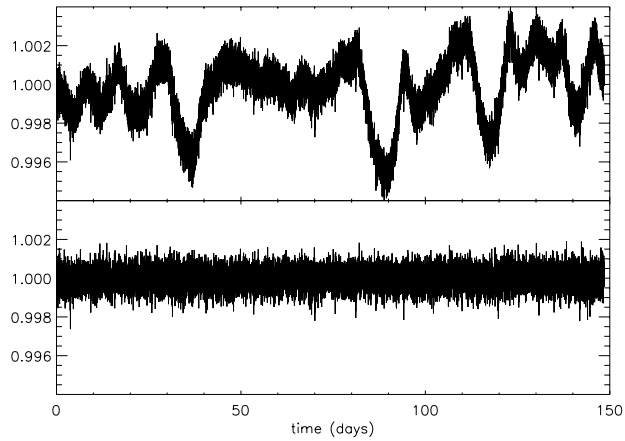


Fig. 5. Light curve 34 before (top) and after (bottom) iterative non-linear filtering with a trial duration of 3.3 hrs (Team 5). The Y-axis represents a relative flux.

Detecting the transits: The box-shaped transit finding algorithm of Aigrain & Irwin (2004) is applied to the filtered light curves. This algorithm, based on likelihood maximisation of a box-shaped, periodic transit model, maximises the transit signal-to-noise ratio $S = \sqrt{N_{\rm tr}} \times \Delta F/\sigma$, where $N_{\rm tr}$ is the number of in-transit points, ΔF the transit depth (which is the mean deviation from the median of the in-transit points) and σ is the robustly estimated scatter. The parameters are the transit duration, period and epoch. Note that the optimal transit depth is fully determined by the light curve and is thus not a free parameter. The maximum multiple and single transit statistics (S_M and S_S respectively) are then saved and plotted for all light curves (see Fig. 6).

Light curves without events form a clump at low S_S and S_M, while those containing significant residual stellar variations form a tail at high S_S, with $S_M \propto S_S$. A threshold of the form $S_M \geq a + b \times S_S$ was therefore used to pick out periodic events, with $a = 1$ (a makes the threshold more stringent at low S_S's) and $b = 1.3$. All events below a similar line with $b = 1.4$ are marked as low-confidence events. All light curves with $S_S \geq 20$ are also included in the candidate lists as potentially containing single deep transits.

The long-term variation filtering and transit search are run for trial transit durations of 3.3, 6.7 and 13.3 hrs, yielding 3 initial lists of 30, 74 and 167 candidates respectively. After examining the corresponding light curves by eye to remove obviously spurious candidates, the final (merged) list contains 31 candidates, of which 6 are low-confidence detections ($S_M \leq 1.0 + 1.4 \times S_S$), 5 are identified as eclipsing binaries due to visible secondary eclipses, 1 as a triple

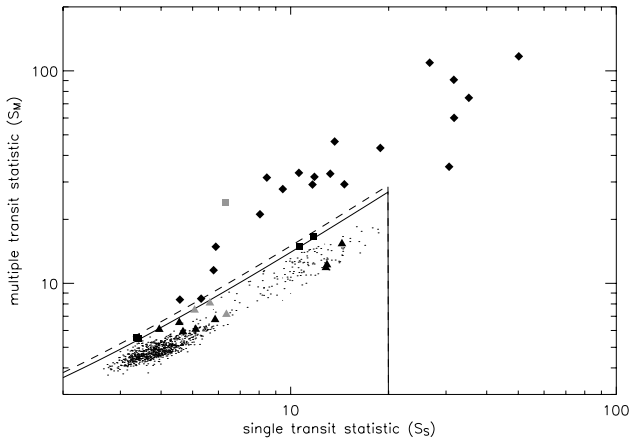


Fig. 6. Candidate selection in the multiple (S_M) versus single (S_S) transit detection statistic plane (trial duration 3.3 hrs). Solid line: detection threshold. Dashed line: low-confidence threshold. Diamonds: correct detections. Squares: false detections (grey: excluded at the light curve examination stage). Triangles: missed detections (grey: detected using another trial duration). (Team 5)

star system and 1 as showing only sinusoidal variations (no transits).

The actual duration of transit candidates is estimated as the full-width at half-minimum of the transits in the phase-folded, filtered light curve. If that differed from the trial duration, the filtering is re-run using the measured duration to obtain a better transit depth measurement (period and epoch were deduced from the transit search itself). Additionally to the transit search, a search for periodic variations with $0.5 \leq P \leq 4$ days is run by sine-fitting, providing improved period estimates for the stellar variables identified by the transit search, and one additional detection of sinusoidal variation.

4. Comparison and analysis

The analysis of the results is performed in two steps: first, an objective comparison of the individual results, and second, a discussion of each subteam on its own performance.

4.1. "Blind" analysis

Tables 1 and 2 give the details of the detection ability of each team for each transit and other contaminating events. From a direct comparison of the individual results we observe that:

- Nine transits were detected by all teams (ID numbers 34, 85, 207, 390, 460, 474, 533, 835, 917). These are clearly validated by 5 independent detections. The measured parameters are very similar, except for the transit duration, whose estimation probably depends on the measurement protocol. Periods are always estimated with a $< 0.1\%$ dispersion around the actual value. A Jupiter-type planet with two transits is detected around a m_V =16 G dwarf star. The smallest planet detected by all teams (ID 390) has a radius of 2.4 R_E, a period of 8 days and orbits a bright G-type dwarf star (m =12).
- Seven transit events were not detected by any team (ID numbers 317, 326, 575, 618, 681, 715). One (ID 715) corresponds to a giant planet in grazing eclipse on a strongly variable, faint and large-radius star. The other non-detected transits correspond to small planets (1.8 to 2.5 R_E) and possibly illustrate the detection limit expected for CoRoT (Fig. 7 and next section).
- Five transits were detected by only some of the teams. ID 168, 537, 613 and 624 were detected by 2 to 4 teams. They correspond to star-planet systems with more than 10 transit events in the total light curve duration. Finally, ID 915, detected by two teams, had a "trick": it is a 58-day period planet around a binary star; only two teams have seen the planet whereas the binary was easily detected by all teams.
- Nine false positives were announced. Teams 2, 4 and 5 detected 1, 3 and 5 false events respectively, while teams 1 and 3 did not detect any. It never happens that a false event is detected by two independent teams on the same light curve. This probably represents the most remarkable result of this study. This result is very positive as it clearly proves that false alarms are method-dependent. It is probable that using such independent multiple analyses will considerably reduce the false-alarm rate also with real data. Again, this result is not specific to CoRoT.
- Six eclipsing binaries and variable stars were detected by all teams but sometimes wrongly identified as planetary transits when they are background or grazing eclipsing binaries (ID 31, 249, 386, 919, 937, 985). Note that such cases should be identified by spectroscopic or/and photometric ground-based follow-up.
- Three of the contaminating events are not detected. ID 271 and 650 are non-periodic variables, and thus do not affect the transit search. ID 518 is the only eclipsing binary which is never detected, but it has only two shallow transits over the 150-day period (grazing binary with an M-dwarf companion).

Quantitative comparisons on the computing time required do not evidence major discrepancies between the teams; moreover, this was not always the priority of the detection teams to minimize the analysis time, so that a crude comparison is not realistic at this level. Eventhough, no analysis method requires a computing time which is uncompatible with the data flow expected from space transit searches. Also, none of the methods described in this paper is strongly sensitive to the short and frequent gaps in the data due to the SAA. Finally, the strong residual scattered-light noise never limits the detection, whatever method is used, even when an over-correction of the scattered light led to a periodic, negative signal, more easily confused with transit signatures.

The results show that the simple correlation method proposed by Team 1 is already a performant detection tool (22 detected events over 38 inserted). It also appears that teams 3 and 5 have detected significantly more transit events than the three other teams (26 detected events). Team 3, moreover, had no false positive, compared to five false positives for team 5. Team 5 could have included less false positives with a higher threshold (see Figure 6), but the method of team 3 has the additional advantage of a very natural way of setting the threshold (Figure 4 bottom). This points towards a greater robustness of the method used by team 3. It confirms that the BLS algorithm is more sensitive to faint transits, a result which also shows up in the theoretical comparison performed by Tingley (2003) or in the recent re-analysis of the OGLE data (Udalski et al. 2003). The better results of team 3 could also be due to a more efficient detrending technique.

4.2. Derived detection limits of CoRoT

Figure 7 shows the three types of results (5 detections, 1 to 4 detections, 0 detection) against the main parameters that affect the detection sensitivity: transit depth d and number of transits n in the light curve. The non-detected events are all situated below the empirical detection curve $d \simeq 2\ 10^{-3} n^{-1/2}$, except one which corresponds to a difficult case described earlier (ID 715). The detection capability of CoRoT derived from this blind test analysis (where r is the planet radius and R the star radius) are:
- $n = 50$ (period < 3 days): $r > 0.017R$ is detected.
- $n = 15$ (period < 10 days): $r > 0.023R$ is detected.
- $n = 3$ (period < 50 days): $r > 0.034R$ is detected.

This "law" may overestimate the minimum detected size when the number of transits is small. It also does not account for the detrending of systematics, which may have an unpredictable impact on the detection.

Table 3 gives the corresponding values of the minimal detected planet size for four types of parent stars, F0V, G0V, K0V and M0V.

4.3. Lessons learned per team

4.3.1. Team 1

Non-detections: The residual pollution by stellar microvariability may explain some non-detections. Light curve ID 915, where the transit is mixed with a fast eclipsing binary, was missed because such a possibility was not considered. It would however not have been detected since the small event was embedded in a secondary detection peak. It shows one of the detection limits of the method.

False detections: None, due to the low sensitivity limit of the method and to visual elimination steps.

Prospects for further improvements: The periodicity of the transit signal could be used in the detection. The removal of the instrumental noise could be improved with a tool such as Principal-Component Analysis. Filtering the long-period variations would also be useful.

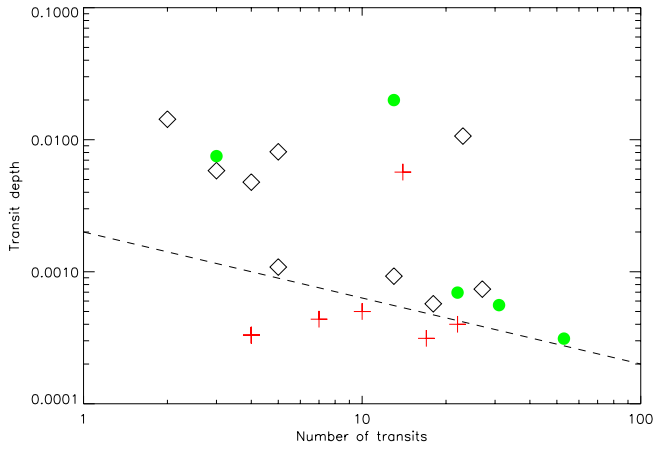


Fig. 7. Depth of the transits versus number of transits. Plus signs show the non-detected events, diamonds show the events detected by five groups independently, and filled circles correspond to 1 to 4 detection occurences. The dashed line thus shows the border of the simulated CoRoT detection limit (proportional to $n^{-1/2}$). The only plus sign above the detection line is a grazing planet on a faint fluctuating star.

4.3.2. Team 2

Non-detections: For most of the non-detections some individual events were detected on a very low confidence level, but most signals were below the detection limit of our routine. To detect these events a search in folded light curves is necessary to improve the S/N ratio of the signals. For ID 168, several transit-shaped events were detected with a medium confidence level, but many were resulting from the variability of the stars, confusing the detection algorithm searching for a periodicity. Consequently the treatment of the variability of the stars and the robustness of the periodicity search has to be improved.

False detections: Only one false detection was made by the team. In light curve ID 213, simulating a faint star, a false transit event was found. This detection had the lowest confidence level of all our detections (3σ).

Prospects for further improvements: A first step would be to search for transits in folded light curves to be able to detect fainter transits in noisier environments. We also plan to test a Fourier analysis and remove frequencies that can be identified as instrumental noise. The deformation of the transit events can be prevented that way. Additionally the light curves of variable stars have to be analysed more carefully to reduce periodic variations that can confuse the detection algorithm.

4.3.3. Team 3

Non-detections; False detections: We have not detected six of the simulated transits, and we had no false detec-

tions. Examining the non-detections reveals that the correct peak appears in the SR for two of them – ID numbers 317 and 575. In order to detect light curve 575 we would have to lower the detection threshold to 5.6, which would have resulted in 86 false detections! The score of light curve 317 was 3.9, which would mean an unrealistic lowering of the threshold. Most of the non-detected (ID numbers 317, 326, 575, 618 and 681) transits corresponded to planets of the smallest radii – less than $0.023\,R_{\odot}$. Light curve 715 was affected by the presence of two different periods and escaped detection. Visual inspection of some of the non-detected transits also suggests that maybe some additional variability still exists after removal of the harmonics, but this effect was not quantified yet.

Prospects for further improvements: The detrending process may benefit significantly from new procedures recently developed for systematic-effect removal (Tamuz et al. 2005; Kovács et al. 2005). This procedure may remove a significant part of the stellar variability, but also some systematic effects that were not modelled in this exercise.

The detection stage may benefit from the correction proposed by Tingley (2003) to the BLS algorithm. In principle, the corrected BLS should be somewhat more powerful in distinguishing between a transit signal and random noise, thus improving the detection ability. Another improvement in the application of the BLS may be related to a better sampling of the frequency space, fine tuning of the algorithm parameters (maximum transit width, bin width, etc.), or better adjustment of the SR function. Finally, one could also make a 2-D search that looks at both the "SDE" and "DDE" parameters of Kovács et al. (2002) and check whether this allows some gain in the detection capability.

4.3.4. Team 4

Non-Detections: A posteriori analyses show that the algorithm cannot detect a planet with radius less than $0.02R_{\odot}$ or when the noise (likely stellar noise) is so strong that the denoising algorithm starts modifying the transit itself.

False detections: Among the detected signals, three of them turn out to be false detection (IDs 701, 703, 983). The case of ID 983 corresponds to a discontinuity of the light curve produced by the stellar variability simulation (sect. 2.2.1). In the other cases (IDs 701, 703), transit features were erroneously identified with the peak sorting method due to a random and unlucky location of the peaks in the convolution curve M. This kind of false detection should however not be specific to our algorithm.Our best results are obtained when the matched filter is associated with a peak selection by Fourier correlation. No false alarm is found in this case, while selection by peak sorting can lead to a number of false alarms due to ambiguities with noise artefacts. Finally, the number of false positives does not change with the detection threshold, which is automatically optimized from an estimate of the noise in the input signal. The threshold thus strongly depends on the quality of the detrending process.

Prospects for further improvement: The method developed in the present exercise can certainly be improved for higher noise level.A new filter based on image processing is presently being tested to improve the detection capacity. It is developed on the same ground as the detrending tool presented in section 3.4.
Another issue is the actual robustness of the algorithms to periodicity changes, due for example to binarity, secondary planets or residual instrumental drift. This question has not been addressed in the present exercise since transit signals were assumed strictly periodic.

4.3.5. Team 5

Non detections; false detections: Three noteworthy points emerge from Fig. 6. First, the tail of small dots with high S_S and $S_M \simeq S_S$ (ie no clear evidence of a periodic signal) represents light curves with residual non-periodic variations. In some cases, these were on too short a timescale to be fully filtered. However, some of the stellar light curves built with the method of Section 2.2.1 contained transit-like features, which are probably artefacts. Second, all the false alarms that escaped removal at the light curve examination stage were low-confidence events. Had the higher threshold of $S_M \geq 1.0+1.4S_S$ been used, there would have been no false alarms. This is the result of a conscious decision to include low-confidence detections,in order to pinpoint the detectability limit. Third, there were 11 missed detections for the 3.3 hr trial duration. Of those, 3 were detected at other trial durations and one was a planet orbiting a binary causing non-periodic events, which we didn't tailor our search for. By lowering the threshold, we could have detected ID 575 and ID 317 at the cost of 1 and 17 additional false alarms respectively. Of the others, 3 produced S_M's close to our algorithm's white-noise limit of ~ 6 (Aigrain & Irwin 2004) and are thus probably beyond the detectability limit of CoRoT. The other 3 were highly variable light curves.

Prospects for further improvements: Future improvements will include refinement of the detrending stages, of the choice of threshold through Monte Carlo simulations, and of the post-detection transit characterisation.

5. Conclusion and prospects

The present paper reports on the first "real size" blind study of a large set of synthetic stellar light curves, by five independent expert teams, to detect planetary transits. Different transit detection methods were tested, ranging from ad-hoc procedures built from scratch to seasoned algorithms used in existing ground-based surveys (OGLE and BEST) and for future space missions (CoRoT and Kepler). Together, they offer a global view of most possible approaches to light curve detrending and transit detection

algorithms. The results show that specialised algorithms can detect transit signals down to the noise limit. It turns out that rather simple procedures can identify most detectable transit signals, but that the additional effort put into refined detection algorithms is really worthwhile to recover the shallower detectable transits – potentially the most interesting ones.

The results also show that false detections may not be a major difficulty when various detection methods are applied, since no false event was ever detected twice independently in the simulation. Also, one method (harmonic-fitting filtering plus BLS detection) does not suffer from any false detections on the synthetic sample. We note that stellar micro-variability limits the transit detection only when its standard deviation is larger than 0.5 % and its main frequency is around 0.1 day^{-1}. In most cases, stellar micro-variability such as simulated here (section 2.2) is not the main limitation, mostly because the fluctuation frequencies are not in the domain of the transit duration, and the amplitude is usually low. This result compares well with the conclusions of Jenkins (2003) and are important in the context of space transit detection missions. Of course, this is true only as far as activity models based on the solar case correctly describe other stars. In the next few years, space astero-seismology missions may provide better constraints on stellar micro-variability on timescales of a few hours.

The present study shows that the detrending method is almost as important for the detection of faint transits as the detection algorithm itself. Precise detrending process can cancel off almost all the variability and reflected light contamination. On the other hand, artefacts of the detrending can cause spurious transit detections. The relative importance of detrending and detection could be quantified by coupling the detrending and detection phases between the five algorithms. This was not attempted in the present study.

The processing of real data will suffer more systematic effects than those introduced in this exercise. In space, these are due to temperature cycles, pointing jitter or scattered light gradients along the detector. In ground-based transit surveys, systematics are mainly due to fluctuations in the Earth atmosphere. Future work will include a comparative study of the gain brought by the correction of systematics using comparison stars, such as recently proposed by Tamuz et al. (2005) and Kovács et al. (2005).

The characterisation of transits (shape, radius ratio, orbital inclination, etc) requires an entirely different set of analysis tools and no particular insight was obtained about it from the detection simulation – apart from confirming that eclipsing binaries can easily be confused with planetary transits.

Some of the algorithms used here focus on the detection of individual transits as well as strictly periodic signals. Detection of not strictly periodic transit signal is an issue that was not considered here. In most realistic cases (two planets, circumbinary planet), the transits will be very nearly periodic and the algorithms for periodic signals will probably be able to detect them.

Among the algorithms studied here, at least two have reached "maturity" for monochromatic light curves without a priori information. Continuation of this study could consider the inclusion of more information: e.g. chromatic light curves (CoRoT), colour or spectroscopic information about the target star. It could also include other instrumental contents (Kepler, Eddington) and a refinement of stellar micro-variability in the frequency-amplitude parameter zone where it may mimic transit features.

The 999 light curves produced and a table with used parameters are made available to the community by request to the authors for testing and improving other detection algorithms.

Another by-product of our blind comparison of detecting transits in light curves simulated as CoRoT data, is a refined estimate of the detection limitation of this instrument to come: a 3-day $1.1R_E$ planet around an M0 dwarf star would probably be detected. CoRoT would also detect the transits of a planet like μ *Arae d*, the 14.5-Earth mass planet with 9.55-day period recently discovered in radial-velocity surveys (Santos et al. 2004), if it is larger than 2.7 R_E, i.e. with a density up to that of terrestrial planets.

Acknowledgements. We are grateful to the CoRoT PI Annie Baglin and to the whole CoRoT/Exoplanet Working Group for their support and fruitful discussions on this exercise. S.Z. wishes to acknowledge support by the European RTN "The Origin of Planetary Systems" (PLANETS, contract number HPRN-CT-2002-00308) in the form of a fellowship. Finally, we express our acknowledgements to the anonymous referee for his/her detailed reading and many interesting suggestions.

References

Aigrain, S. & Favata, F. 2002, A&A, 395, 625
Aigrain, S., Favata, F., & Gilmore, G. 2004, A&A, 414, 1139
Aigrain, S. & Irwin, M. 2004, MNRAS, 350, 331
Allen, C. 2000, "Astrophysical Quantities", A.N. Cox editions
Andersen, B. N., Leifsen, T., & Toutain, T. 1994, Sol. Phys., 152, 247
Andreasen, J. 1988, A&A, 196, 159
Auvergne, M., Boisnard, L., & Buey, J.-T. 2003, SPIE, 4853, 170
Baglin, A. 2003, Adv. Sp. Res., 31, 345
Barban, C., Goupil, M. J., & Van't Veer-Menneret, C. 2003, A&A, 405, 1095
Boisnard, L. & Auvergne, M. 2004, in 55 th IAF Congress, October 4-7 2004 Vancouver, IAC-04-Q.1.01
Bordé, P., Rouan, D., & Léger, A. 2003, A&A, 405, 1137
Borucki, W., Koch, D., Boss, A., et al. 2004, in Second Eddington Workshop: Stellar structure and habitable planet finding. Edited by F. Favata, S. Aigrain and A. Wilson. ESA SP-538, p. 177
Bouchy, F., Pont, F., & Santos, N. 2004, A&A, 421, L13

Brown, T. 2003, ApJ, 593, 125
Chapman, G. A., Cookson, A., & Dobias, J. 1997, ApJ, 482, 541
Deeg, H. 1999, Universal Transit Modeler, *www.iac.es/galeria/hdeeg/idl_hans_lib/utm/*
Defaÿ, C. 2001, PhD Thesis
Defaÿ, C., Deleuil, M., & Barge, P. 2001, A&A, 365, 330
Doyle, L. R., Deeg, H. J., Kozhevnikov, V. P., et al. 2000, ApJ, 535, 338
Favata, F. 2004, in Second Eddington Workshop: Stellar structure and habitable planet finding. Edited by F. Favata, S. Aigrain and A. Wilson. ESA SP-538, p. 3
Guis, V. & Barge, P. 2004, PASP, accepted
Jenkins, J., Doyle, L., & Cullers, D. 1996, Icarus, 119, 244
Konacki, M., Torres, G., Jha, S., & Sasselov, D. 2003, Nature, 421, 507
Konacki, M., Torres, G., & Sasselov, D. 2005, submitted, ApJ
Kovács, G., Bakos, G., & Noyes, R. 2005, MNRAS, 356, 557
Kovács, G., Zucker, S., & Mazeh, T. 2002, A&A, 391, 369
Lanza, A. F., Rodonò, M., Pagano, I., Barge, P., & Llebaria, A. 2003, A&A, 403, 1135
Lanza, A. F., Rodonò, M., & Pagano, I. 2004, A&A, 425, 707
Llebaria, A., Guterman, P., & Ollivier, M. 2003, SPIE, 5170, 155
Mandel, K. & Agol, E. 2002, ApJ, 580, L171
Messina, S., Pizzolato, N., Guinan, E., & Rodonò, M. 2003, A&A, 410, 671
Pont, F., Bouchy, F., Queloz, D., et al. 2004, A&A, 426, L15
Press, W., Teukolsky, S., Vetterling, W. T., & Flannery, B. P. 1993, Numerical Recipes in C (2nd ed.; Cambridge Univ. Press)
Santos, N., Bouchy, F., Mayor, M., Pepe, F., & Queloz, D. 2004, A&A, 426, L19
Tamuz, O., Mazeh, T., & Zucker, S. 2005, MNRAS, 356, 1466
Tingley, B. 2003, A&A, 408, L5
Udalski, A., Paczynski, B., Zebrun, K., et al. 2002a, Acta Astron., 52, 1
Udalski, A., Pietrzynski, G., Szymanski, M., et al. 2003, Acta Astron., 53, 133
Udalski, A., Szymanski, M. K., Kubiak, M., et al. 2004, Acta Astronomica, 54, 313
Udalski, A., Zebrun, K., Szymanski, M., et al. 2002b, Acta Astron., 52, 115
Walker, G., Matthews, J., Kuschnig, R., et al. 2003, PASP, 115, 1023
Wichmann, R. 1998, Nightfall *www.lsw.uni – heidelberg.de/users/rwichman/Nightfall.html*

Table 1. The characteristics of the transits that were inserted in the light curves: the star radius R (in solar radius units), the stellar limb darkening coefficient (LD), the planet radius r, the orbital period in days, the system inclination in degrees, the semi-major axis a, the star magnitude, the final standard deviation of the light curve in percents, and some comments. The detection flag shows a series of + and − signs, corresponding to each team, respectively from 1 to 5; + means a positive detection (for team 1 in position 1, etc...), − means that the event is missed.

ID	R ($R_\odot$)	LD	r ($R_\odot$)	period (days)	inc (deg)	a ($R_\odot$)	m	Std Dev. %	Comment	Detection Flag
34	0.92	0.6	0.025	5.52	88.8	12.77	13	0.17		+ + + + +
85	1.1	0.4	0.099	26.4	88.9	37.876	15	0.75		+ + + + +
168	0.92	0.5	0.13	11.5	87.4	20.827	15	0.71		− − + + +
207	0.92	0.5	0.11	88.4	90.0	79.89	16	1.42		+ + + + +
317	1.1	0.6	0.02	33.8	89.5	44.66	12	0.09		− − − − −
326	0.85	0.6	0.017	6.8	89.9	13.9	14	0.40		− − − − −
390	0.92	0.6	0.022	8.0	89.	16.35	12	0.07		+ + + + +
460	1.1	0.3	0.076	32.9	89.52	23.49	15	0.82		+ + + + +
474	0.92	0.6	0.028	11.34	89.	20.63	13	0.18		+ + + + +
533	0.92	0.7	0.095	6.4	90.0	7.89	16	1.54		+ + + + +
537	0.85	0.6	0.015	2.78	89.9	7.68	12	0.09		− − + − +
575	0.85	0.6	0.019	15.9	90.0	24.57	14	0.40		− − − − −
613	1.1	0.6	0.026	4.8	89.4	12.16	14	0.29		+ − + − +
618	1.3	0.6	0.023	8.48	89.	19.55	12	0.09		− − − − −
624	1.1	0.6	0.029	6.7	89.8	15.18	14	0.29		+ − + + +
681	1.1	0.6	0.023	19.8	89.6	31.27	13	0.20		− − − − −
715	1.3	0.3	0.098	10.1	86.4	21.96	15	0.75	Planet 1	− − − − −
			0.07	63.8	89.7	75.0			Planet 2	− − − − −
835	1.1	0.4	0.084	42.6	89.3	52.10	15	0.74		+ + + + +
915	1.5	0.25	0.13	58.32	89.9	70.0	15	0.74	Planet	− + + − −
		0.3	1.1	2.9	86.0	11.4			Binary	+ + + + +
917	0.85	0.6	0.028	30.4	89.7	37.8	13	0.18		+ + + + +

Table 2. Table of contaminating events which were introduced in the light curves: magnitude, event type ("BEB" stands for background eclipsing binaries, "GrB" stands for grazing binaries), period and relative flux (contribution of the background star to the total flux), standard deviation of the final light curve. Detection flag: detection and correct identification (+), wrong identification (i), no detection (−), for each team from 1 to 5. References: UTM (Deeg, 1999, UTM), Nightfall (Wichmann, 1998, W98), (Mandel & Agol, 2002, MA), AAVSO (American Association of Variable Star Observers), Andreasen (1988) (A88).

ID	mv	Event type	period (days)	relative flux	Std Dev. %	Reference	Detection Flag
31	14	BEB	24.7	0.03	0.38	UTM	+ + + + +
131	14	δ Cepheid	5.86	-	0.40	A88	− − − + −
249	16	triple star	3.9	-	1.03	UTM	+ + + + +
259	15	GrB	1.4132	-	0.82	W98	+ − + + +
271	15	Z Cam	-	-	0.82	AAVSO	− − − − −
384	15	β cephei	0.2835	0.001	0.81	AAVSO	+ − − + +
386	15	GrB	17.1	-	0.66	UTM	i i i i i
486	15	BEB	2.4128	0.01	0.66	UTM	− − i − +
518	15	GrB	78.3	-	0.82	MA	− − − − −
553	15	δ scuti	0.07342	0.003	0.66	AAVSO	− − + + +
599	15	GrB	1.874	-	0.75	W98	+ − + i +
650	14	semi regular	-	-	0.37	AAVSO	− − − − −
809	15	GrB	3.2	-	0.75	W98	i − + + +
919	16	GrB	13.2	-	1.02	UTM	i + + i i
937	15	BEB	8.452	0.01	0.81	UTM	i i i i i
985	15	BEB	5.19	0.01	0.71	UTM	i i i i i

Table 3. Minimum planet radius for F0V, G0V, K0V and M0V stars, in unit of Earth radius, corresponding to the empirical detection curve estimated by the blind test, which possibly overestimates the minimal radius of the detected planets at the longest periods. The star radii are from Allen (2000), i.e. 1.5, 1.1, 0.85 and 0.6 solar radius, respectively.

Period	F0V	G0V	K0V	M0V
50-day	5.6	4.0	3.2	2.2
10-day	3.75	2.75	2.1	1.5
3-day	2.8	2.0	1.6	1.1

Bibliographie

Aigrain, S. & Favata, F., *Bayesian detection of planetary transits. A modified version of the Gregory-Loredo method for Bayesian periodic signal detection.* 2002, A&A, 395, 625

Aigrain, S., Gilmore, G., Favata, F., & Carpano, S., *The Frequency Content of the VIRGO/SoHO Light Curves : Implications for Planetary Transit Detection from Space.* 2003, in ASP Conf. Ser. 294 : Scientific Frontiers in Research on Extrasolar Planets, 441–444

Aigrain, S. & Irwin, M., *Practical planet prospecting.* 2004, Monthly Notices of the Royal Astronomical Society, 350, 331

Baglin, A., Auvergne, M., Catala, C., Michel, E., & COROT Team, *Asteroseismology with the space mission CoRoT : photometric performances targets and mission profile.* 2001, in ESA SP-464 : SOHO 10/GONG 2000 Workshop : Helio- and Asteroseismology at the Dawn of the Millennium, 395–398

Barge, P. & Sommeria, J., *Did planet formation begin inside persistent gaseous vortices ?* 1995, A&A, 295, L1

Bordé, P., Rouan, D., & Léger, A., *Exo-planet detection with the CoRoT space mission. I. A multi-transit detection criterion.* 2001, Academie des Sciences Paris Comptes Rendus Serie Physique Astrophysique, 7, 1049

Bordé, P., Rouan, D., & Léger, A., *Exoplanet detection capability of the CoRoT space mission.* 2003, A &A, 405, 1137

Bordé, P., *Détection et Caractérisation de Planètes Extrasolaires par Photométrie Visible et Interférométrie Infrarouge à très haute Précision* . 2003, PhD thesis, Observatoire de Paris

Chauvin, G., Lagrange, A.-M., Dumas, C., et al., *A giant planet candidate near a young brown dwarf. Direct VLT/NACO observations using IR wavefront sensing.* 2004, A&A, 425, L29

Claret, A., *Non-linear limb-darkening law for LTE models (Claret, 2000).* 2000, VizieR Online Data Catalog, 336, 31081

Debray, B., *Photométrie Stellaire dans les Champs Encombrés Pour l'Etude des Galaxies Proches.* 1982, PhD thesis, Universite de droit d'économie et des sciences d'Aix Marseille

Deeg, H. J., Doyle, L. R., Kozhevnikov, V. P., et al., *Near-term detectability of terrestrial extrasolar planets : TEP network observations of CM Draconis.* 1998, A&A, 338, 479

Defaÿ, C., *Traitement du Signal pour la Détection des Transits planétaires : Application à la Mission Spatiale CoRoT.* 2001, PhD thesis, Laboratoire d'Astrophysique Spatiale

Defaÿ, C., Deleuil, M., & Barge, P., *A Bayesian method for the detection of planetary transits.* 2001, Astronomy and Astrophysics, 365, 330

Deleuil, M. & et al., *Detection of Earth-Sized Planets with the CoRoT Space Mission.* 1997, in ASP Conf. Ser. 119 : Planets Beyond the Solar System and the Next Generation of Space Missions, 259–+

Duren, R. M., Dragon, K., Gunter, S. Z., et al., *Systems engineering for the Kepler Mission : a search for terrestrial planets.* 2004, in Optimizing Scientific Return for Astronomy through Information Technologies. Edited by Quinn, Peter J. ; Bridger, Alan. Proceedings of the SPIE, Volume 5497, pp. 16-27 (2004)., 16–27

Epstein, G., Adda, M., Auvergne, M., et al., *CoRoT instrument : Constraints and solutions.* 2000, in The Third MONS Workshop : Science Preparation and Target Selection, 157–+

Everitt & S., B. 2001, Cluster Analysis (Arnold Publications)

Gregory, P. C. & Loredo, T. J., *A new method for the detection of a periodic signal of unknown shape and period.* 1992, ApJ, 398, 146

Guterman, P., Barge, P., Llebaria, A., & Quentin, C., *Improving transit detection with collective light curves information.* 2005, in Techniques and Instrumentation for Detection of Exoplanets II. Edited by Coulter, Daniel R. Proceedings of the SPIE, Volume 5905, pp. 155-166 (2005).

Harvey, J. W., Duvall, T. L., Jefferies, S. M., & Pomerantz, M. A., *Chromospheric Oscillations and the Background Spectrum.* 1993, in ASP Conf. Ser. 42 : GONG 1992. Seismic Investigation of the Sun and Stars, 111–+

Henry, G. W., Marcy, G., Butler, R. P., & Vogt, S. S., *HD 209458*. 1999, IAU Circ., 7307, 1

Jain, A. & Dubes, R. 1988, Algorithms for Clustering Data (Englewood Cli s, NJ : Prentice Hall)

Jenkins, J. M., *The Impact of Solar-like Variability on the Detectability of Transiting Terrestrial Planets*. 2002, ApJ, 575, 493

Jenkins, J. M., Doyle, L. R., & Cullers, D. K., *A Matched Filter Method for Ground-Based Sub-Noise Detection of Terrestrial Extrasolar Planets in Eclipsing Binaries : Application to CM Draconis*. 1996, Icarus, 119, 244

Kay, S. 1998, Fundamentals of Statistical Signal Processing, Volume 2 : Detection Theory (Prentice Hall)

Kocher, P., Jaffe, J., & Jun, B., *Differential Power Analysis*. 1999, Lecture Notes in Computer Science, 1666, 388

Konacki, M., *An extrasolar giant planet in a close triple-star system*. 2005, Nature, 436, 230

Kovács, G., Zucker, S., & Mazeh, T., *A box-fitting algorithm in the search for periodic transits*. 2002, Astronomy and Astrophysics, 391, 369

Lanza, A. F., Rodonò, M., Pagano, I., Barge, P., & Llebaria, A., *Modelling the rotational modulation of the Sun as a star*. 2003, A&A, 403, 1135

Lecavelier des Etangs, A., Vidal-Madjar, A., McConnell, J. C., & Hébrard, G., *Atmospheric escape from hot Jupiters*. 2004, A&A, 418, L1

Llebaria, A., Auvergne, M., & Perruchot, S., *Design of polychromatic PSFs for the CoRoT experiment*. 2004, in Optical Design and Engineering. Edited by Mazuray, Laurent ; Rogers, Philip J. ; Wartmann, Rolf. Proceedings of the SPIE, Volume 5249, pp. 175-181 (2004)., 175–181

Llebaria, A., Guterman, P., & Ollivier, M., *Photometric masking methods and predicted performances for the CoRoT exoplanetary mission*. 2003, in Techniques and Instrumentation for Detection of Exoplanets. Edited by Coulter, Daniel R. Proceedings of the SPIE, Volume 5170, pp. 155-166 (2003)., 155–166

Llebaria, A., Vuillemin, A., Guterman, P., & Barge, P., *Designing photometric patterns for exoplanet transit search on board CoRoT*. 2002, in Highly Innovative Space Telescope Concepts Edited by Howard A. MacEwen. Proceedings of the SPIE, Volume 4849, pp. 112-123 2002., 112–123

Mandel, K. & Agol, E., *Analytic Light Curves for Planetary Transit Searches.* 2002, The Astrophysical Journal, 580, L171

Marcy, G. W. & Butler, R. P., *First three planets.* 1996, in Proc. SPIE Vol. 2704, p. 46-49, The Search for Extraterrestrial Intelligence (SETI) in the Optical Spectrum II, Stuart A. Kingsley ; Guillermo A. Lemarchand ; Eds., 46–49

Mayor, M. & Queloz, D., *A Jupiter-Mass Companion to a Solar-Type Star.* 1995, Nature, 378, 355

Moutou, C., Pont, F., Barge, P., et al., *Comparative blind test of five planetary transit detection algorithms on realistic synthetic light curves.* 2005, A&A, 437, 355

Neuhäuser, R., Guenther, E. W., Wuchterl, G., et al., *Evidence for a co-moving sub-stellar companion of GQ Lup.* 2005, A&A, 435, L13

Pepe, F., Mayor, M., Rupprecht, G., et al., *HARPS : ESO's coming planet searcher. Chasing exoplanets with the La Silla 3.6-m telescope.* 2002, The Messenger, 110, 9

Pickles, A. J., *A Stellar Spectral Flux Library : 1150 - 25000 A (Pickles 1998).* 1998, VizieR Online Data Catalog, 611, 863

Press, H., Teukolsky, A., Vetterling, T., & Flannery, P. 1997, Numerical Recipes in C, The Art of Scientific Programming (Cambridge University Press)

Robin, A. & Creze, M., *Stellar populations in the Milky Way - Comparisons of a synthetic model with star counts in nine fields.* 1986, A&AS, 64, 53

Rouan, D., Baglin, A., Barge, P., et al., *Searching for exosolar planets with the CoRoT space mission.* 1999, Physics and Chemistry of the Earth C, 24, 567

Rouan, D., Baglin, A., Copet, E., et al., *The Exosolar Planets Program of the CoRoT satellite.* 2000, Earth Moon and Planets, 81, 79

Schneider. 2005, *The Extrasolar Planets Encyclopaedia*, http ://www.obspm.fr/planets

Schneider, J., *The study of extrasolar planets : methods of detection, first discoveries and future perspectives.* 1999, Academie des Sciences Comptes Rendus Serie Mecanique Physique Chimie Sciences de la Terre et de l Univers, 327, 621

site internet. 2004, *CoRoT Week 6*, http ://www.ias.u-psud.fr/medoc/cw6/index.php ?action=presentlist

site internet. 2005a, *CoRoT : CNES Web Site*, http ://smsc.cnes.fr/COROT/

site internet. 2005b, *CoRoT : CNES Web Site*, http ://corot.oamp.fr/

site internet. 2005c, *Kepler : NASA Web Site*, http ://www.kepler.arc.nasa.gov/

site internet. 2005d, *Telescope Issac Newton, La Palma, Espagne*, http ://www.ing.iac.es/Astronomy/telescopes/int/index.html

Soderblom, D. R., *Rotational studies of late-type stars. II - Ages of solar-type stars and the rotational history of the sun*. 1983, ApJS, 53, 1

Steller, M., Heihsler, J., Ottacher, H., & Weiss, W. 2002, *From stars to habitable planets, the austrian contribution to the CoRoT mission*

Tamuz, O., Mazeh, T., & Zucker, S., *Correcting systematic effects in a large set of photometric light curves*. 2005, Monthly Notices of the Royal Astronomical Society, 356, 1466

Tingley, B., *Improvements to existing transit detection algorithms and their comparison*. 2003a, Astronomy and Astrophysics, 408, L5

Tingley, B., *A rigorous comparison of different planet detection algorithms*. 2003b, Astronomy and Astrophysics, 403, 329

Torres, G., Konacki, M., Sasselov, D. D., & Jha, S., *The transiting planet OGLE-TR-56b*. 2003, American Astronomical Society Meeting Abstracts, 203,

Udalski, A., Szymanski, M., Kaluzny, J., et al., *The optical gravitational lensing experiment. Discovery of the first candidate microlensing event in the direction of the Galactic Bulge*. 1993, Acta Astronomica, 43, 289

Udalski, A., Szymanski, M., Kaluzny, J., Kubiak, M., & Mateo, M., *The Optical Gravitational Lensing Experiment*. 1992, Acta Astronomica, 42, 253

Vidal-Madjar, A., Désert, J.-M., Lecavelier des Etangs, A., et al., *Detection of Oxygen and Carbon in the Hydrodynamically Escaping Atmosphere of the Extrasolar Planet HD 209458b*. 2004, ApJL, 604, L69

Wolszczan, A., *Confirmation of Earth Mass Planets Orbiting the Millisecond Pulsar PSR :B1257+12*. 1994, Science, 264, 538

Résumé

Le mini-satellite Corot lancé en 2006 utilisera la méthode des transits : Une exoplanète signe son passage devant l'étoile par une brève baisse de flux inférieure au millième. La stabilité et continuité de mesure seront assurées sur 150 jours pour 60.000 étoiles afin d'augmenter le nombre de configurations favorables. La photométrie d'ouverture intègre chaque flux dans un masque de lecture adapté aux multiples bruits, limité à 250 formes différentes pour 12.000 cibles. J'ai étudié des méthodes autorisant cette réduction sans perte notable de signal à bruit. Le tri efficace de masques aléatoires s'avère la plus satisfaisante.

Pour la détection, j'ai développé une méthode qui rehausse le contraste des transits en éliminant les composantes des effets collectifs et de certains artefacts. Après détection temporelle les systématiques sont identifiées, même celles d'origine et de poids inconnus. On fait émerger de nouvelles détections en jaugeant la dispersion autour de ces composantes.

Abstract

The Corot mini-satellite to be launched in 2006 uses the transits method : An exoplanet signs its crossing in front of its star by a short drop in light curve with amplitude smaller than 1 per 1000. The stability and continuity of measurement will be ensured over 150 days for 60.000 stars in order to increase the number of favorable configurations. The aperture photometry integrates each star flux inside a reading mask adapted to all noises, but limited to 250 different shapes for 12.000 targets. I studied methods allowing to lead this reduction keeping low S/N loss. The sorting of random masks appears to be the most efficient.

For detection, I developed a method to raise the contrast of transits by eliminating components of collective effects and of some artifacts. After temporal detection the systematics are identified, even those of unknown origin and weights. Then new detections arise by assessing dispersion around these components.